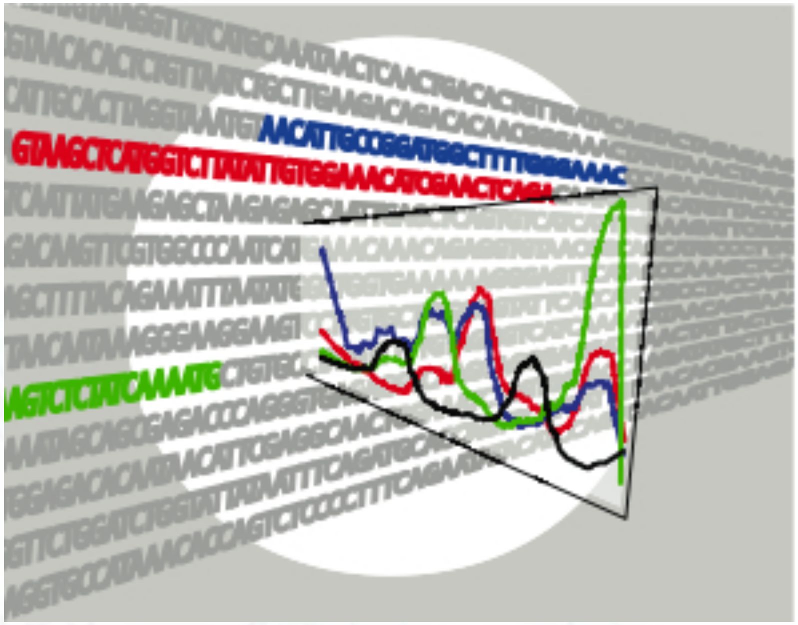
POZNAN SUMMER SCHOOL OF BIOINFORMATICS
DAY 3

06.09.2017

ChIP-SEQ ANALYSIS

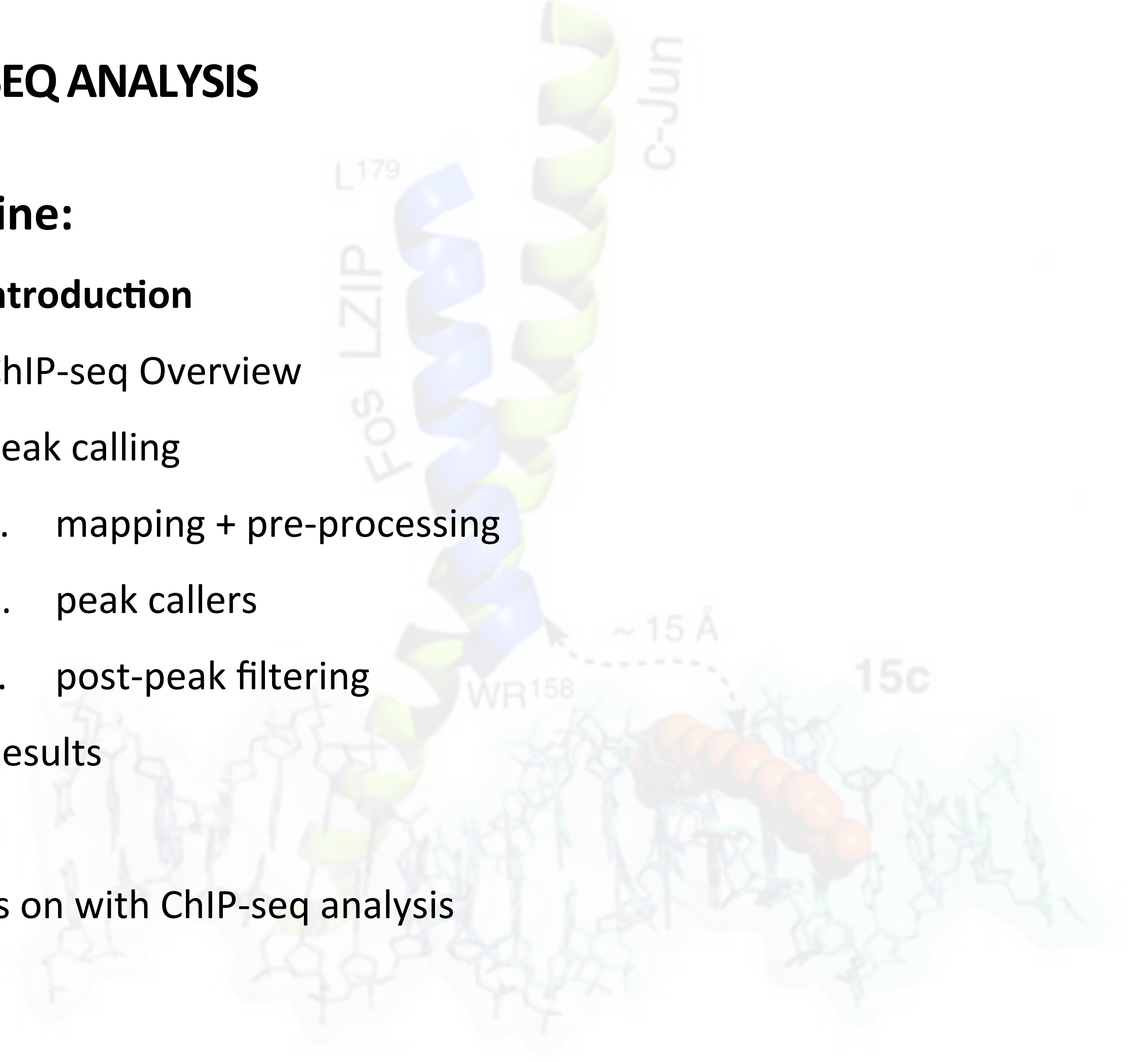
Rebecca Worsley Hunt

rebecca.worsleyhunt@mdc-berlin.de



background Fos-Jun dimer: campusvida.usc.es

ChIP-SEQ ANALYSIS



Outline:

1. **Introduction**
2. ChIP-seq Overview
3. Peak calling
    a. mapping + pre-processing
    b. peak callers
    c. post-peak filtering
4. Results

Hands on with ChIP-seq analysis

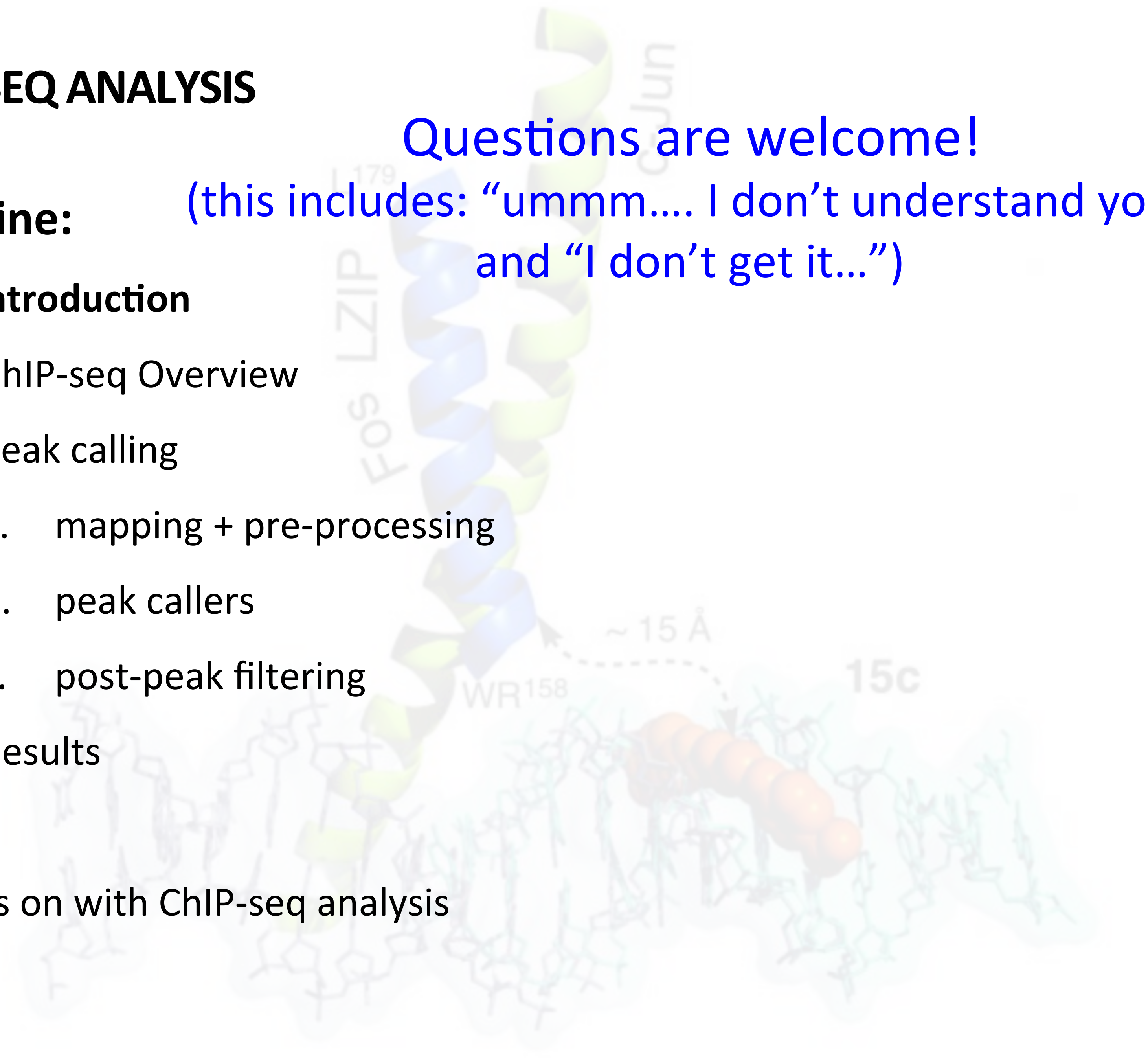
ChIP-SEQ ANALYSIS

Questions are welcome!
(this includes: “ummm…. I don’t understand you” and “I don’t get it…”)

Outline:

1. **Introduction**
2. ChIP-seq Overview
3. Peak calling
   a. mapping + pre-processing
   b. peak callers
   c. post-peak filtering
4. Results

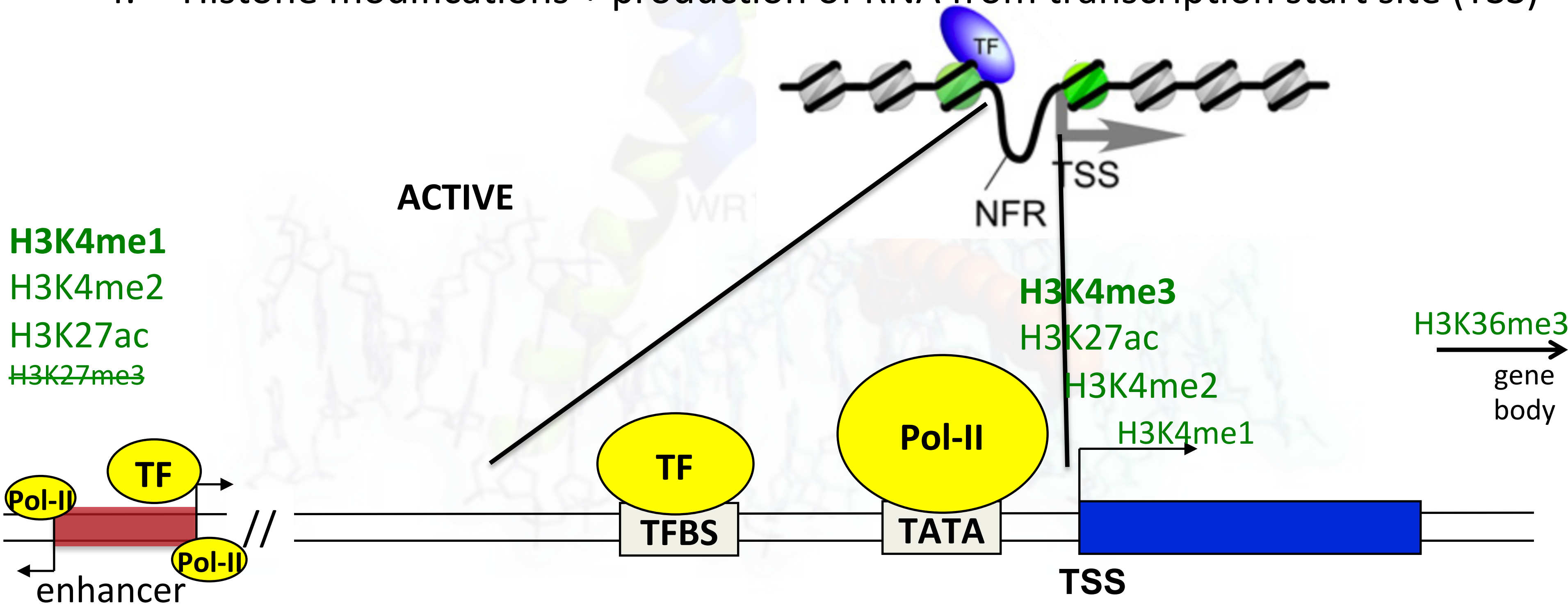
Hands on with ChIP-seq analysis



TRANSCRIPTION INITIATION OVER-SIMPLIFIED

Four-step Process:

1. Histone modification assists chromatin accessibility
2. TFs bind to TFBS (DNA)
3. TFs catalyzes recruitment of polymerase II complex
4. Histone modifications + production of RNA from transcription start site (TSS)



ACTIVE

H3K4me1
H3K4me2
H3K27ac
~~H3K27me3~~

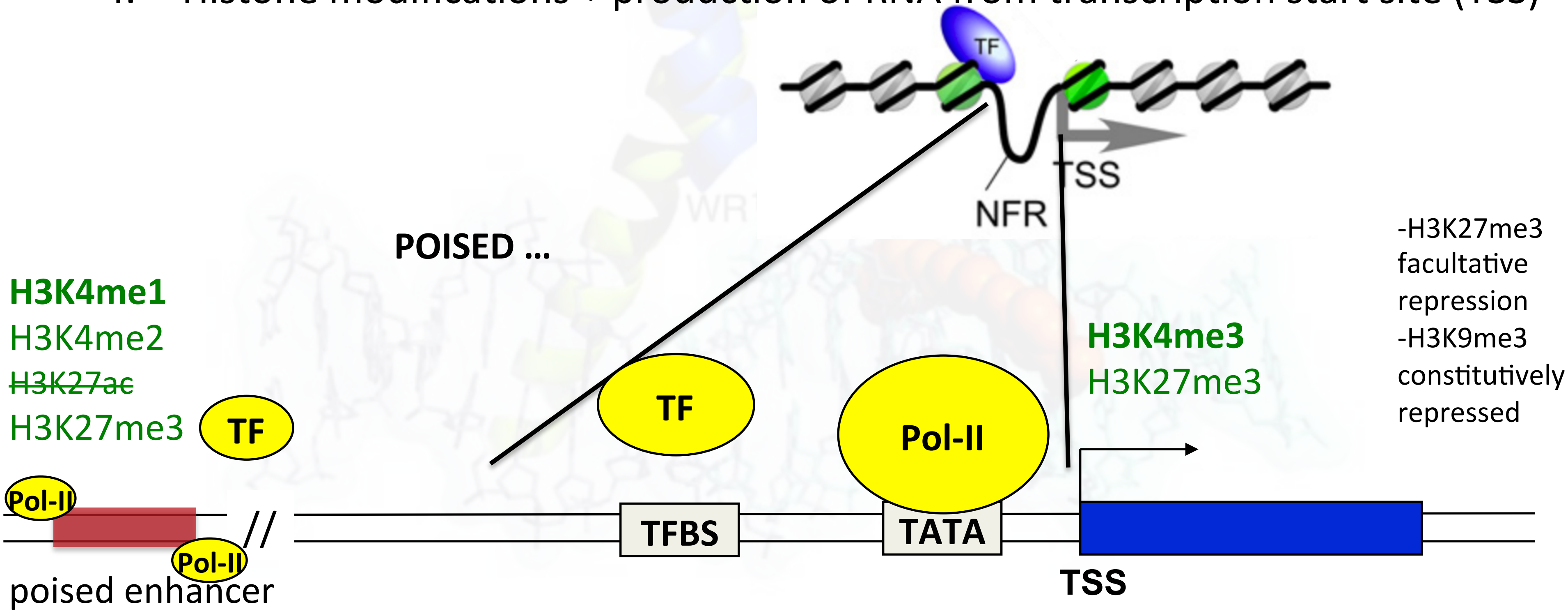
H3K4me3
H3K27ac
H3K4me2
H3K4me1

H3K36me3
gene body

TRANSCRIPTION INITIATION OVER-SIMPLIFIED

Four-step Process:

1. Histone modification assists chromatin accessibility
2. TFs bind to TFBS (DNA)
3. TFs catalyzes recruitment of polymerase II complex
4. Histone modifications + production of RNA from transcription start site (TSS)



POISED …

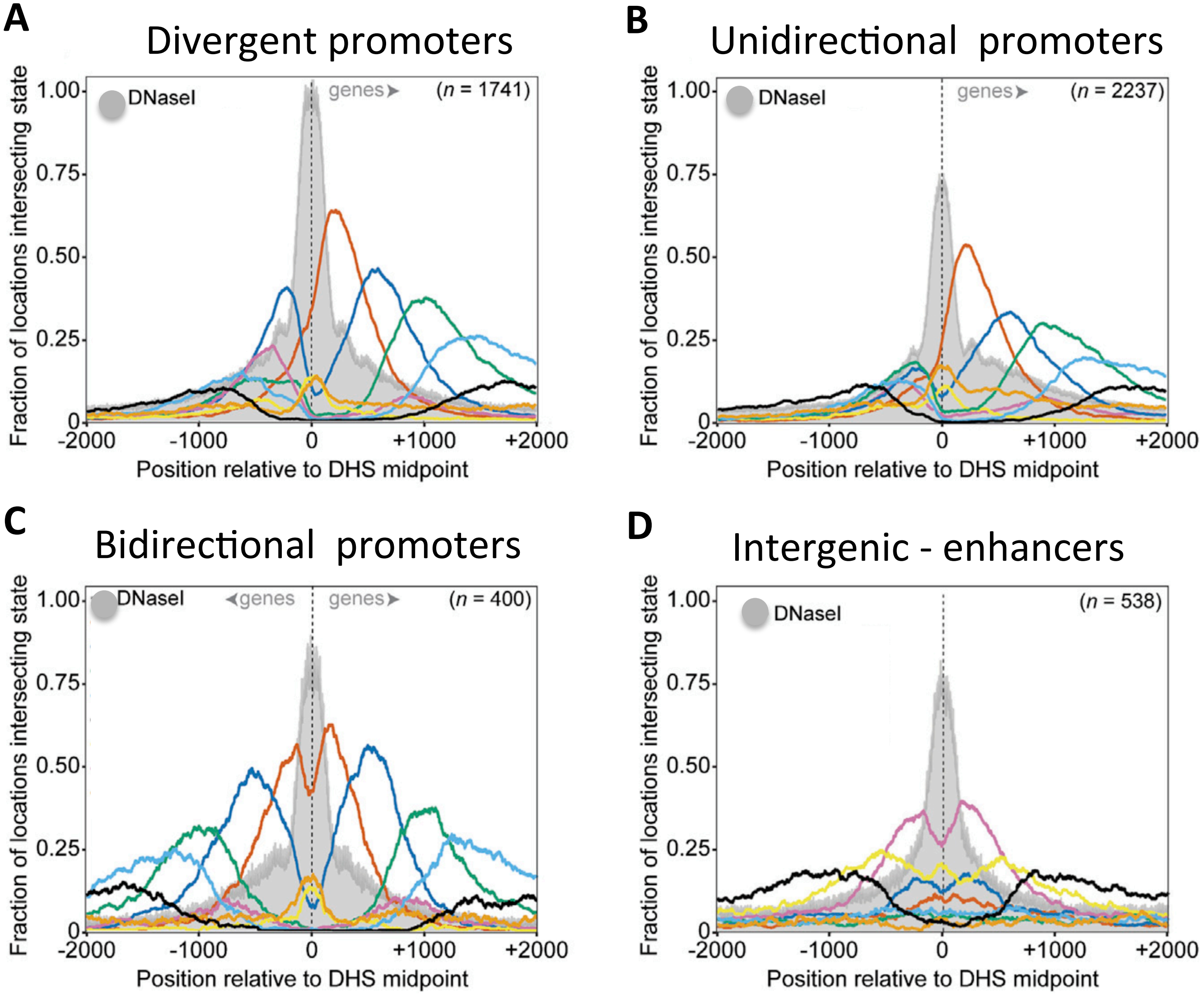
H3K4me1
H3K4me2
~~H3K27ac~~
H3K27me3

poised enhancer

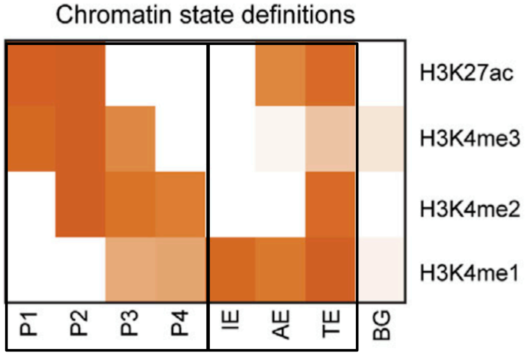
H3K4me3
H3K27me3

-H3K27me3 facultative repression
-H3K9me3 constitutively repressed

NUCLEOSOME-BASED INDICATORS OF TRANSCRIPTIONAL ACTIVITY AT OPEN CHROMATIN AND GENES



4 modifications:

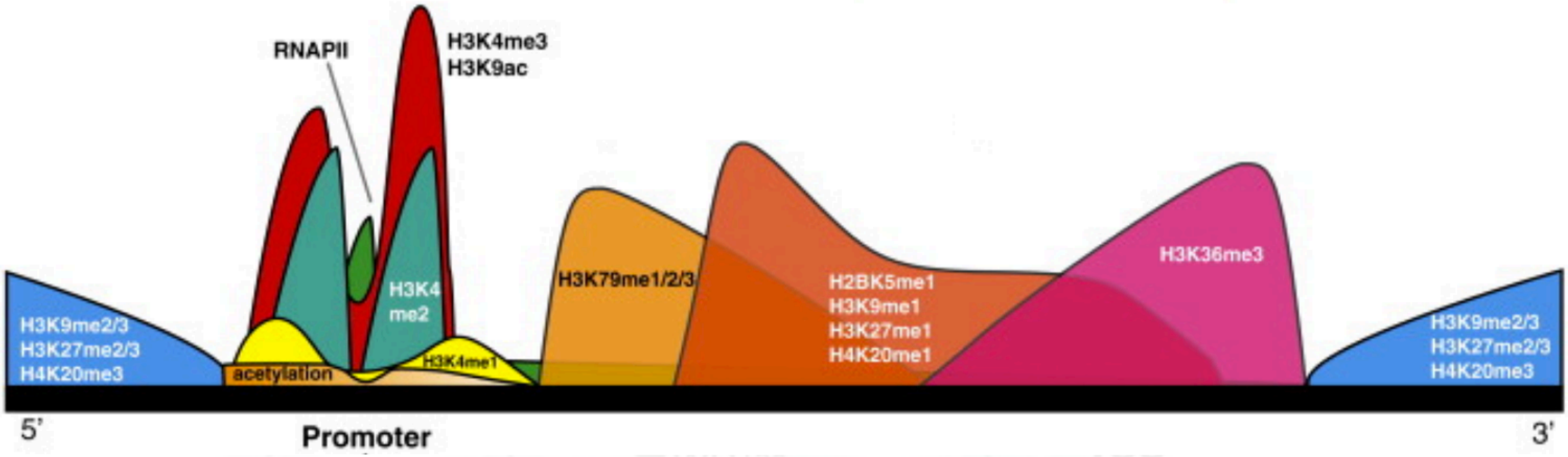


- P1 - Promoter State 1 ●
- P2 - Promoter State 2 ●
- P3 - Promoter State 3 ●
- P4 - Promoter State 4 ●
- IE - Inactive Enhancer ●
- AE - Active Enhancer ●
- TE - Transcribed Enhancer ●
- BG - Background

Duttke S, Lacadie S, et al. Mol Cell. 2015 Feb 19;57(4):674-84

NUCLEOSOME-BASED INDICATORS OF TRANSCRIPTIONAL ACTIVITY AT OPEN CHROMATIN AND GENES

Many modifications



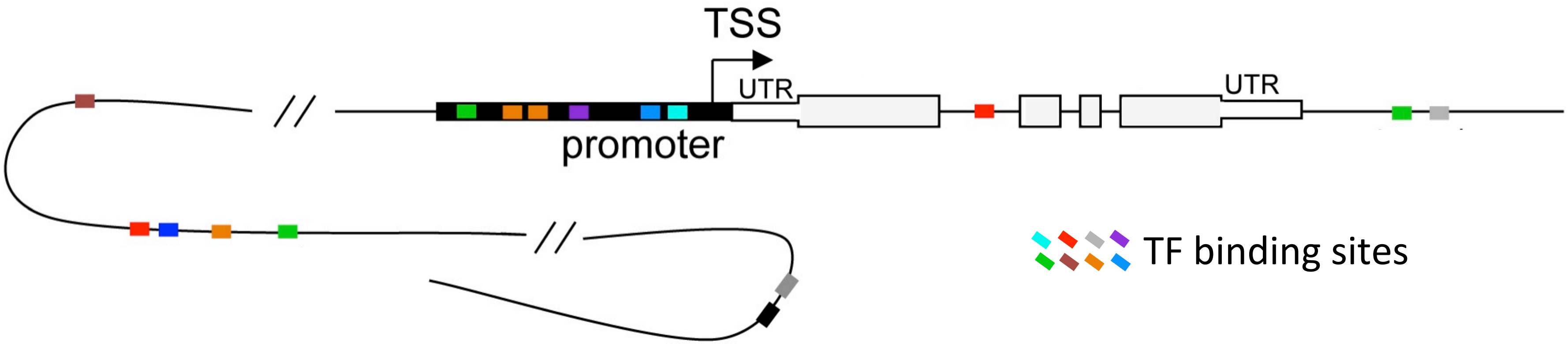
repressive state upstream

active promoter

actively transcribed gene body

Barth and Imhof, Trends Biochem Sci. 2010 Nov;35(11):618-26.

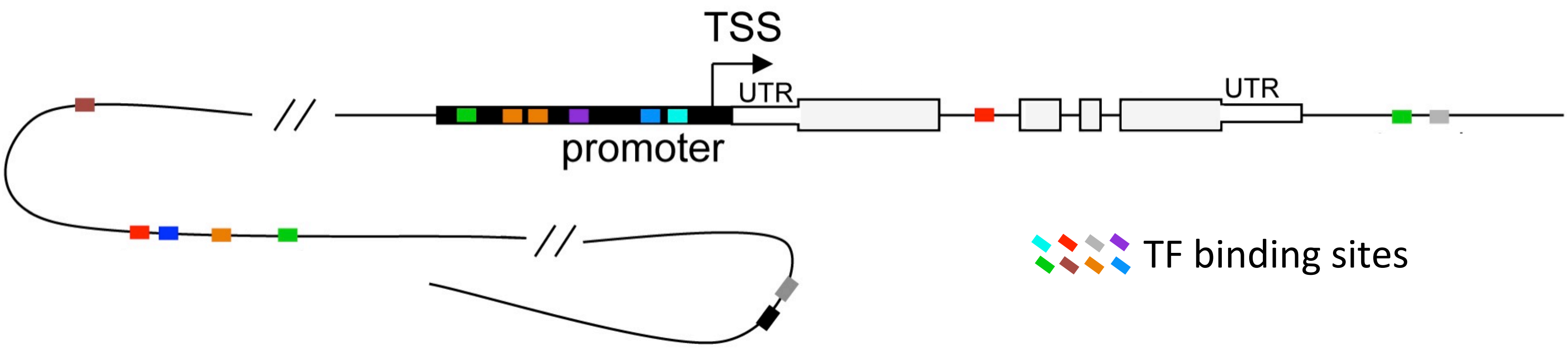
UNDERSTANDING GENE REGULATION – IN HOPE



- Nucleosomes and specific modifications
- Histone modifying proteins -> acetylases, phosphatases, methylases, demethylases etc.
- Transcription factors with/without sequence affinity
- Co-activators
- Polymerases and specific modifications (phosphorylation of Ser5, Ser2 etc)
- Chromatin remodelers (nucleosome organization)
- Cohesins
- Lamin-binding proteins
- Polycomb proteins

… and more

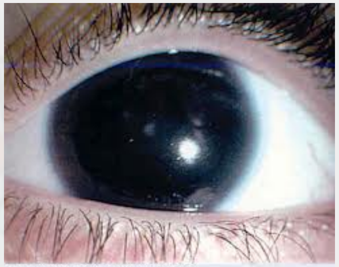
UNDERSTANDING GENE REGULATION – IN HOPE



Hemophilia B
Diabetes
Aniridia
α-thalassemia
Inflammatory bowel disease
Pancreatic agenesis
Otofaciocervical syndrome
…
many cancers and complex diseases

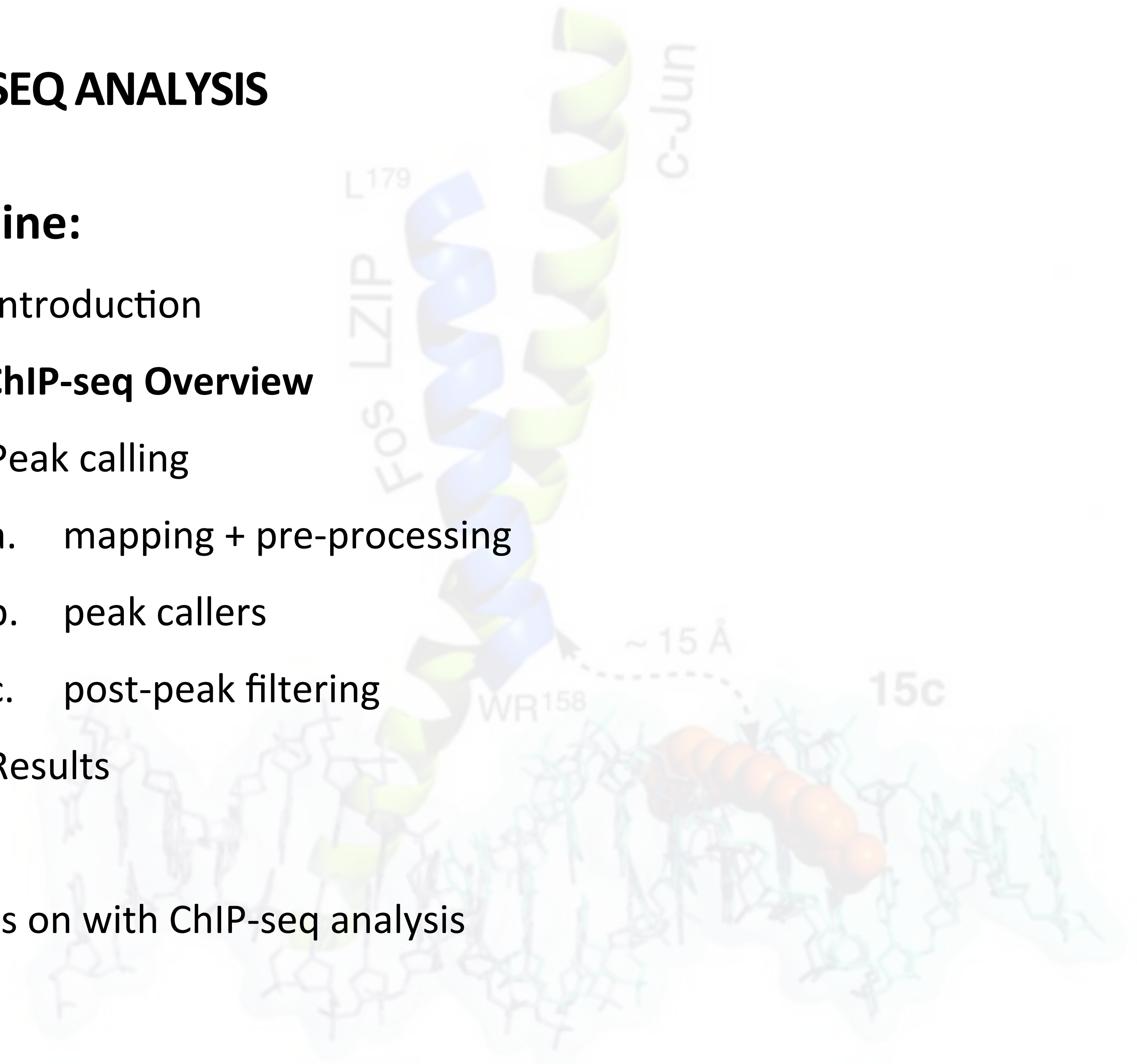
Normal

Aniridia: Pax6



Credit: The Aniridia Foundation International

ChIP-SEQ ANALYSIS

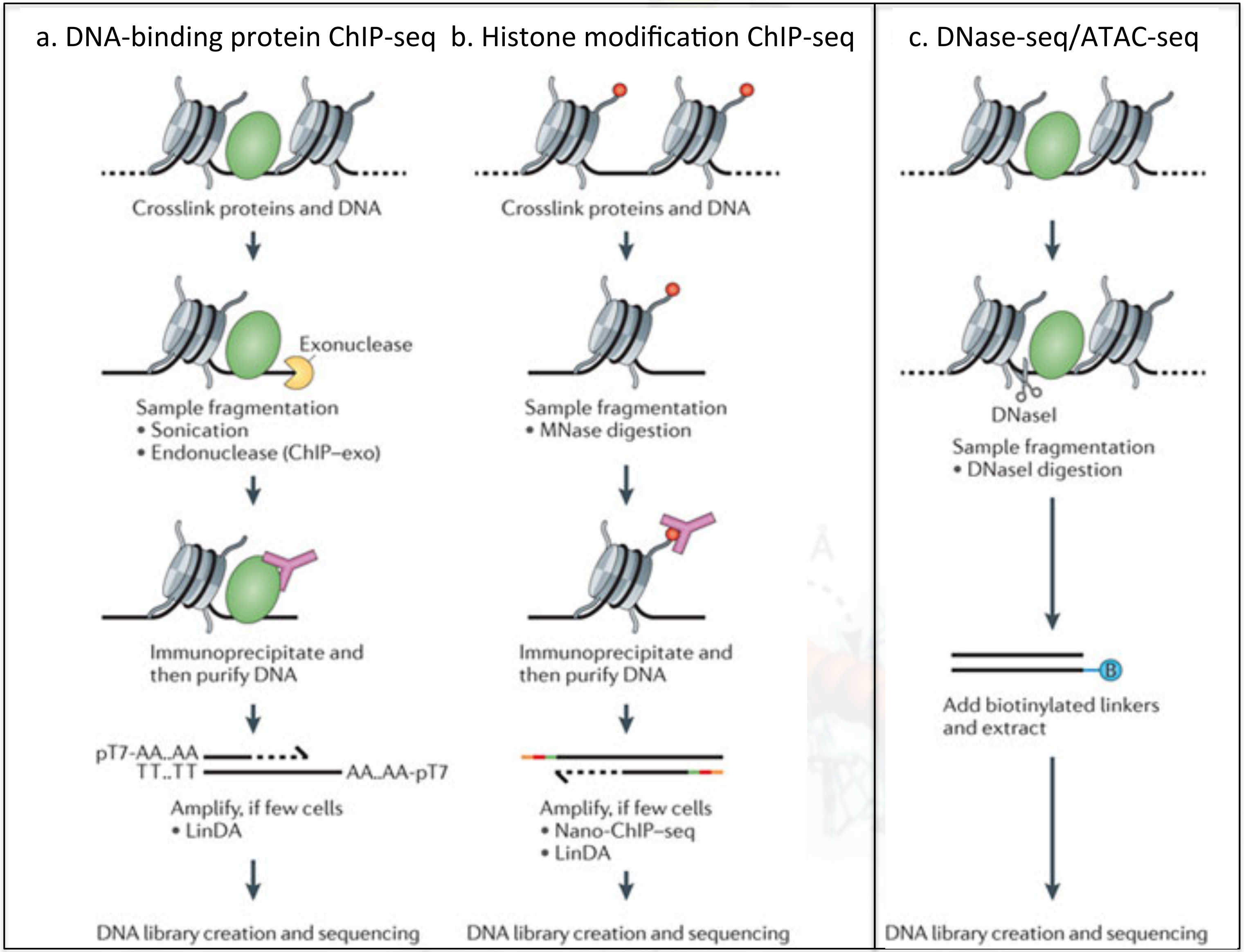


Outline:

1. Introduction
2. **ChIP-seq Overview**
3. Peak calling
   a. mapping + pre-processing
   b. peak callers
   c. post-peak filtering
4. Results

Hands on with ChIP-seq analysis

EXPERIMENTAL PROTOCOLS



a. DNA-binding protein ChIP-seq

Crosslink proteins and DNA

Exonuclease

Sample fragmentation
- Sonication
- Endonuclease (ChIP–exo)

Immunoprecipitate and then purify DNA

pT7-AA..AA
TT..TT
AA..AA-pT7

Amplify, if few cells
- LinDA

DNA library creation and sequencing

b. Histone modification ChIP-seq

Crosslink proteins and DNA

Sample fragmentation
- MNase digestion

Immunoprecipitate and then purify DNA

Amplify, if few cells
- Nano-ChIP–seq
- LinDA

DNA library creation and sequencing

c. DNase-seq/ATAC-seq

DNaseI

Sample fragmentation
- DNaseI digestion

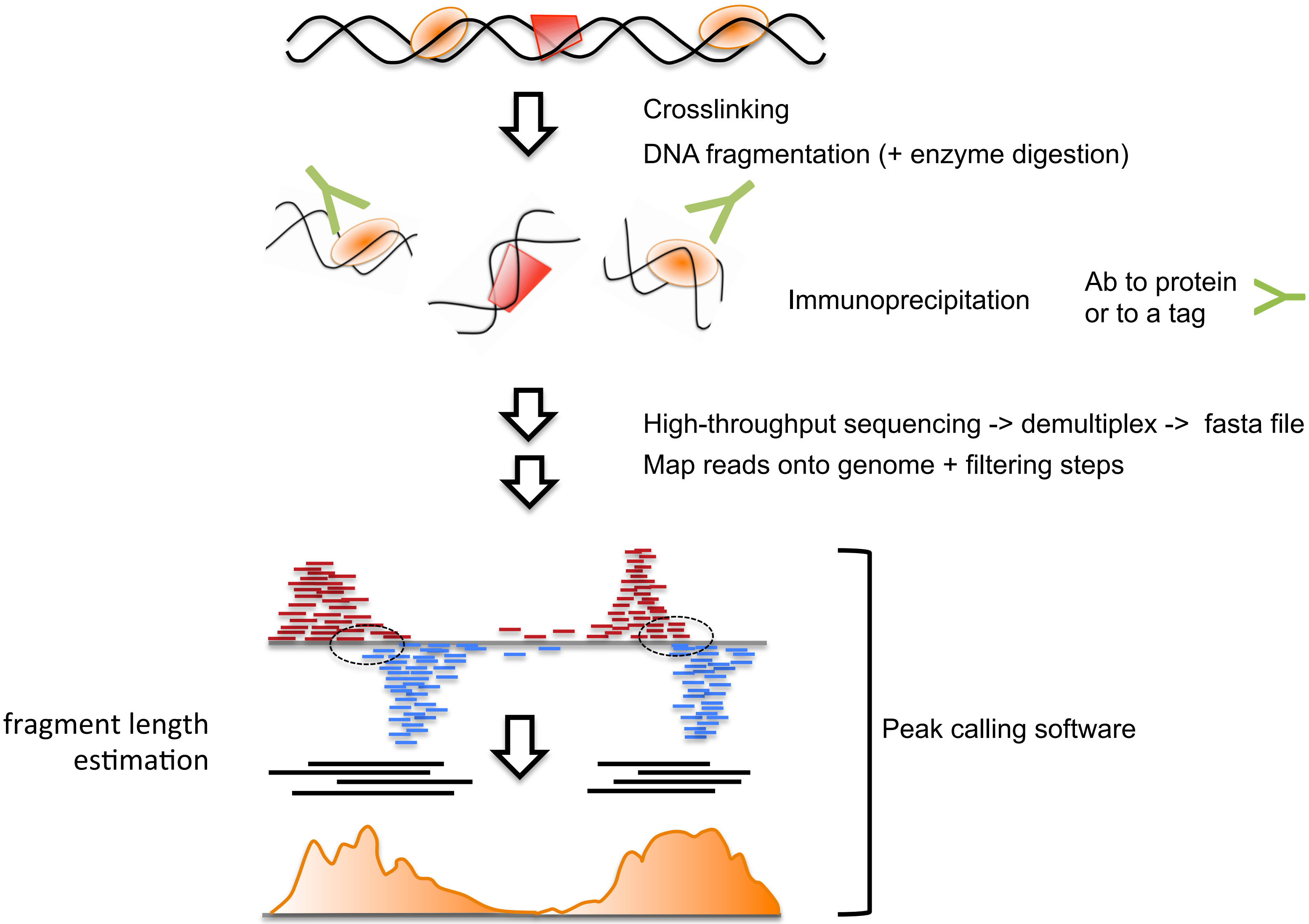
B

Add biotinylated linkers and extract

DNA library creation and sequencing

Furey, Nat Rev Genet 2012

CHIP-SEQ EXPERIMENTS ENRICH FOR REGIONS BOUND BY A PROTEIN OF INTEREST



Crosslinking

DNA fragmentation (+ enzyme digestion)

Immunoprecipitation

Ab to protein
or to a tag

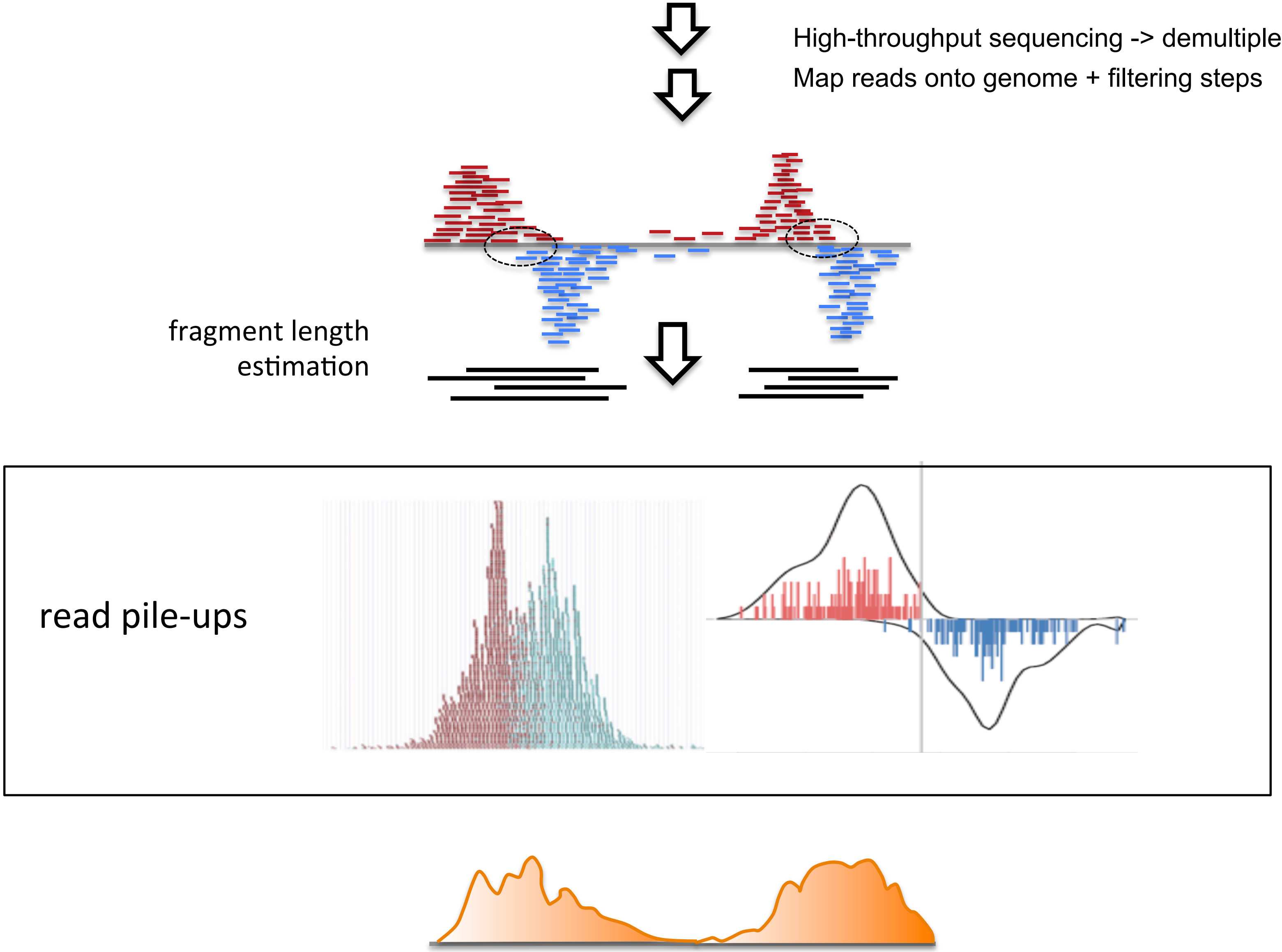
High-throughput sequencing -> demultiplex -> fasta file

Map reads onto genome + filtering steps

fragment length
estimation

Peak calling software

CHIP-SEQ EXPERIMENTS ENRICH FOR REGIONS BOUND BY A PROTEIN OF INTEREST



High-throughput sequencing -> demultiplex -> fasta file

Map reads onto genome + filtering steps

fragment length estimation

read pile-ups

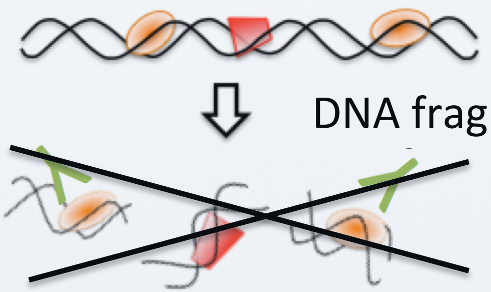
ChIP-SEQ CONTROLS

0. No control
not a good
idea

1. non-induced
condition

no protocol
change

2. Input
(sheared DNA)



crosslink

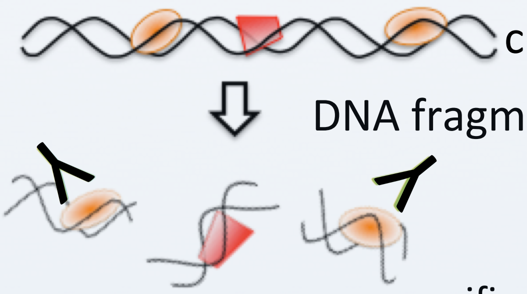
DNA fragmentation

empty
column

~~Immunoprecipitation~~

DNA shearing tendencies

3. IgG



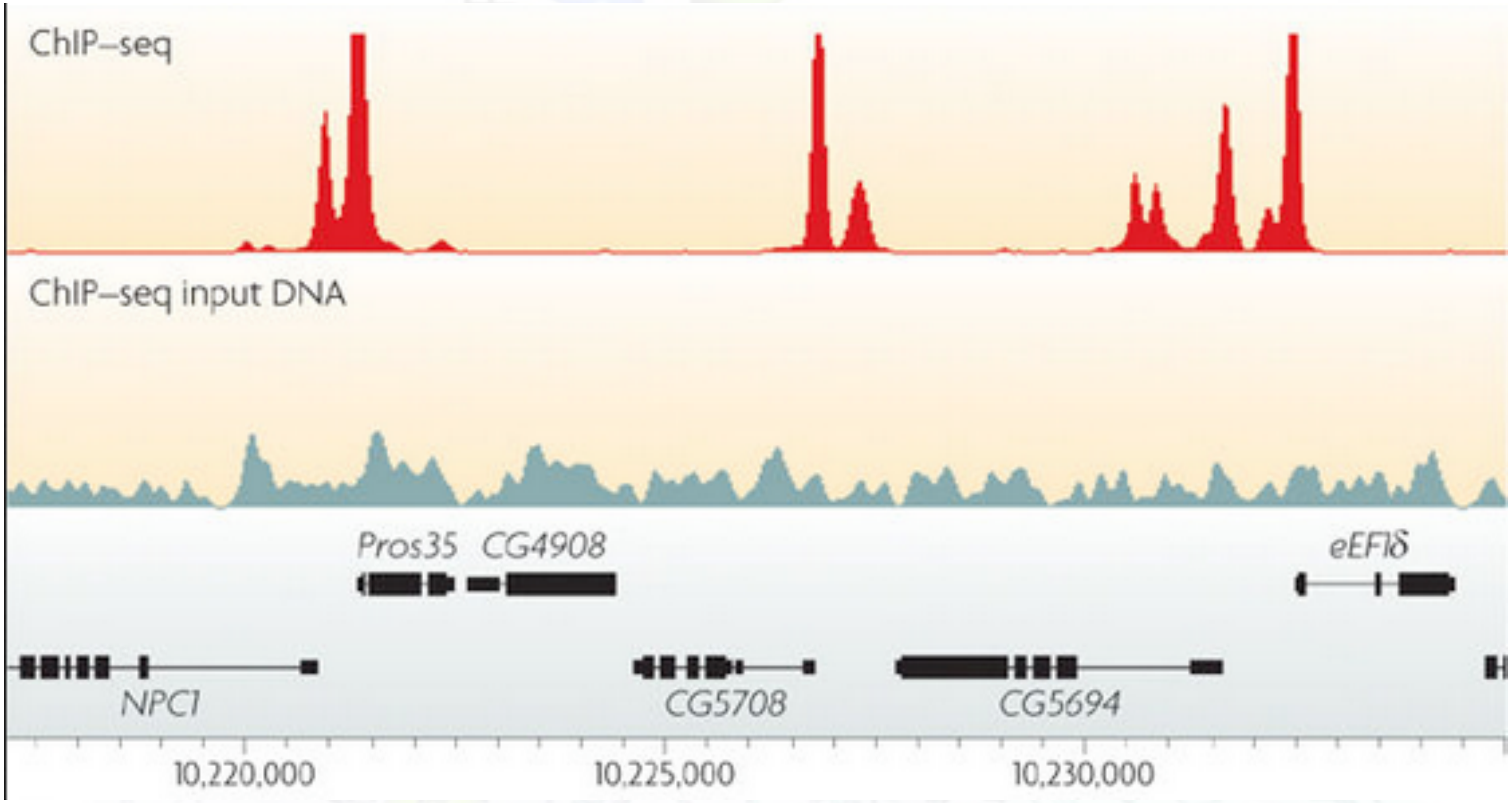
crosslink

DNA fragmentation

non-specific
Immunoprecipitation

DNA “stickiness” to non-specific Ab

CHIP-SEQ EXPERIMENTS ENRICH FOR REGIONS BOUND BY A PROTEIN OF INTEREST



Sims et al Nature Reviews Genetics 15, 121–132 (2014)

DATA RESOURCES

1. You – if many samples, sequence only one or two first!

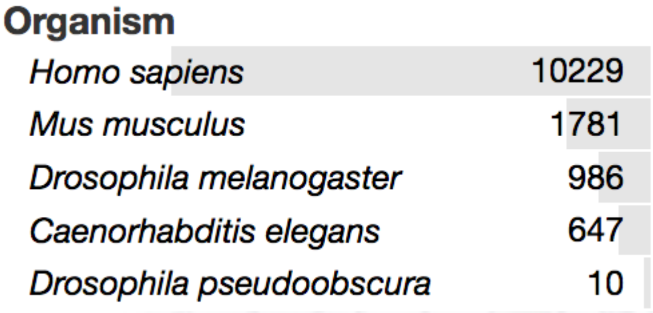
2. Databases

   e.g. UCSC, GEO/ArrayExpress

3. Consortium websites

   e.g. ENCODE, modENCODE, Roadmap Epigenomics (healthy base line)

   https://www.encodeproject.org/matrix/?type=Experiment Experiment Matrix



Organism

Homo sapiens	10229
Mus musculus	1781
Drosophila melanogaster	986
Caenorhabditis elegans	647
Drosophila pseudoobscura	10

Can choose:

- organism
- Sample type e.g. tissue, immortalized cell line, primary cell
- Organ e.g. brain, muscle, liver
- Project
- Genome assembly

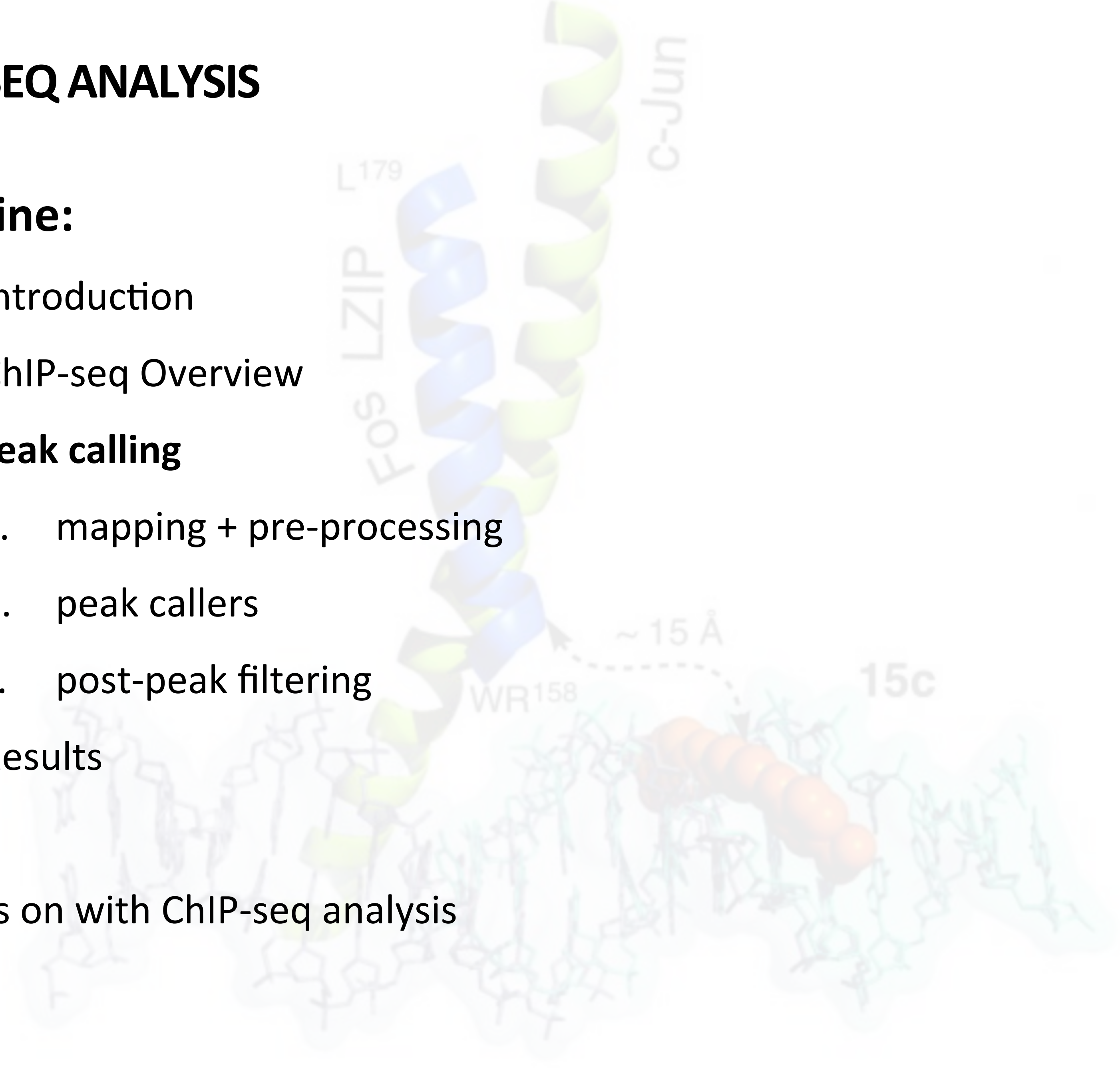
etc

ChIP-SEQ ANALYSIS

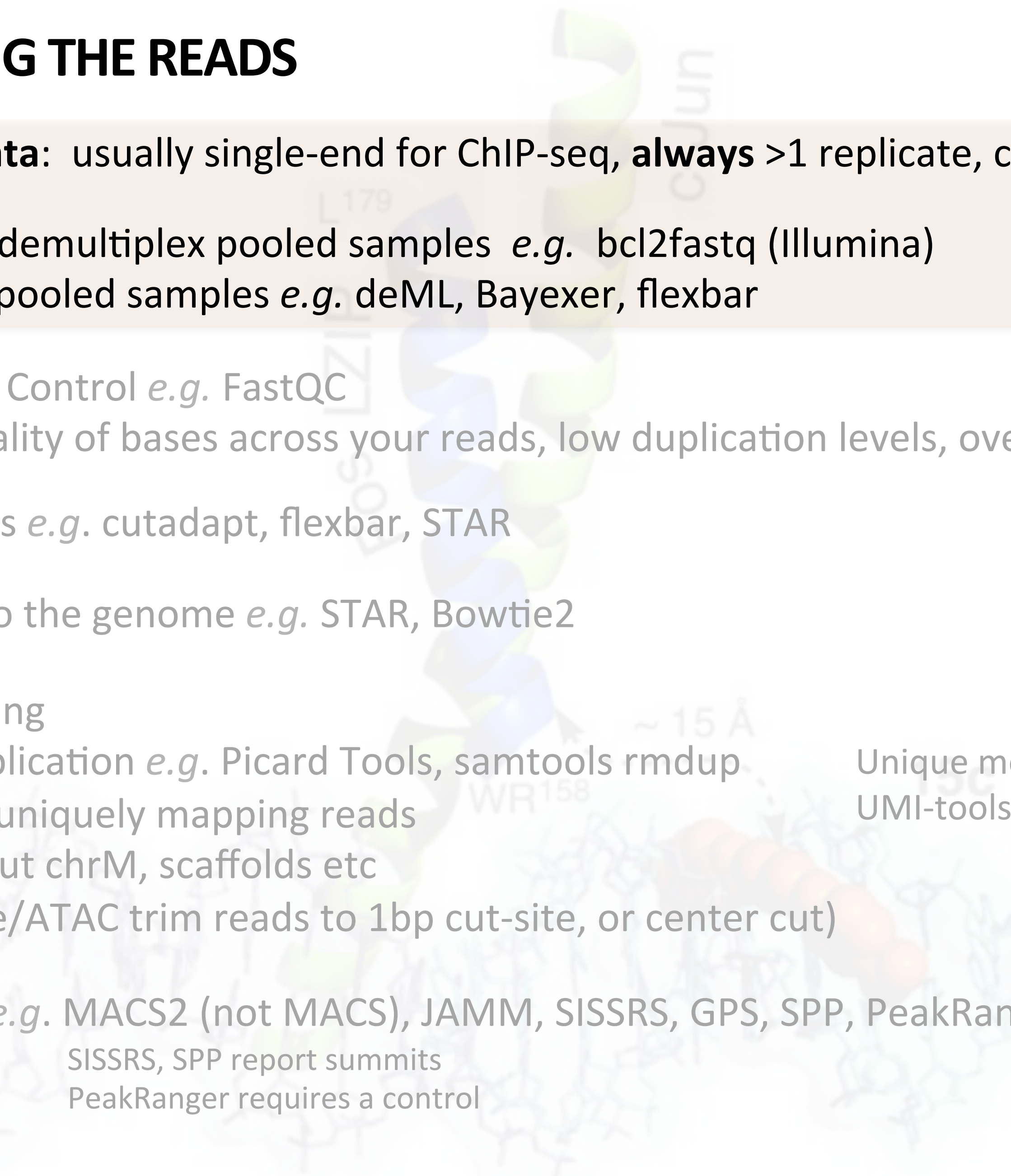
Outline:

1. Introduction
2. ChIP-seq Overview
3. **Peak calling**
    a. mapping + pre-processing
    b. peak callers
    c. post-peak filtering
4. Results

Hands on with ChIP-seq analysis



PROCESSING THE READS



Sequence Data: usually single-end for ChIP-seq, **always** >1 replicate, controls

Convert and demultiplex pooled samples *e.g.* bcl2fastq (Illumina)
Demultiplex pooled samples *e.g.* deML, Bayexer, flexbar

Read Quality Control *e.g.* FastQC
high quality of bases across your reads, low duplication levels, over-represented seqs.

Trim adapters *e.g*. cutadapt, flexbar, STAR

Align reads to the genome *e.g.* STAR, Bowtie2

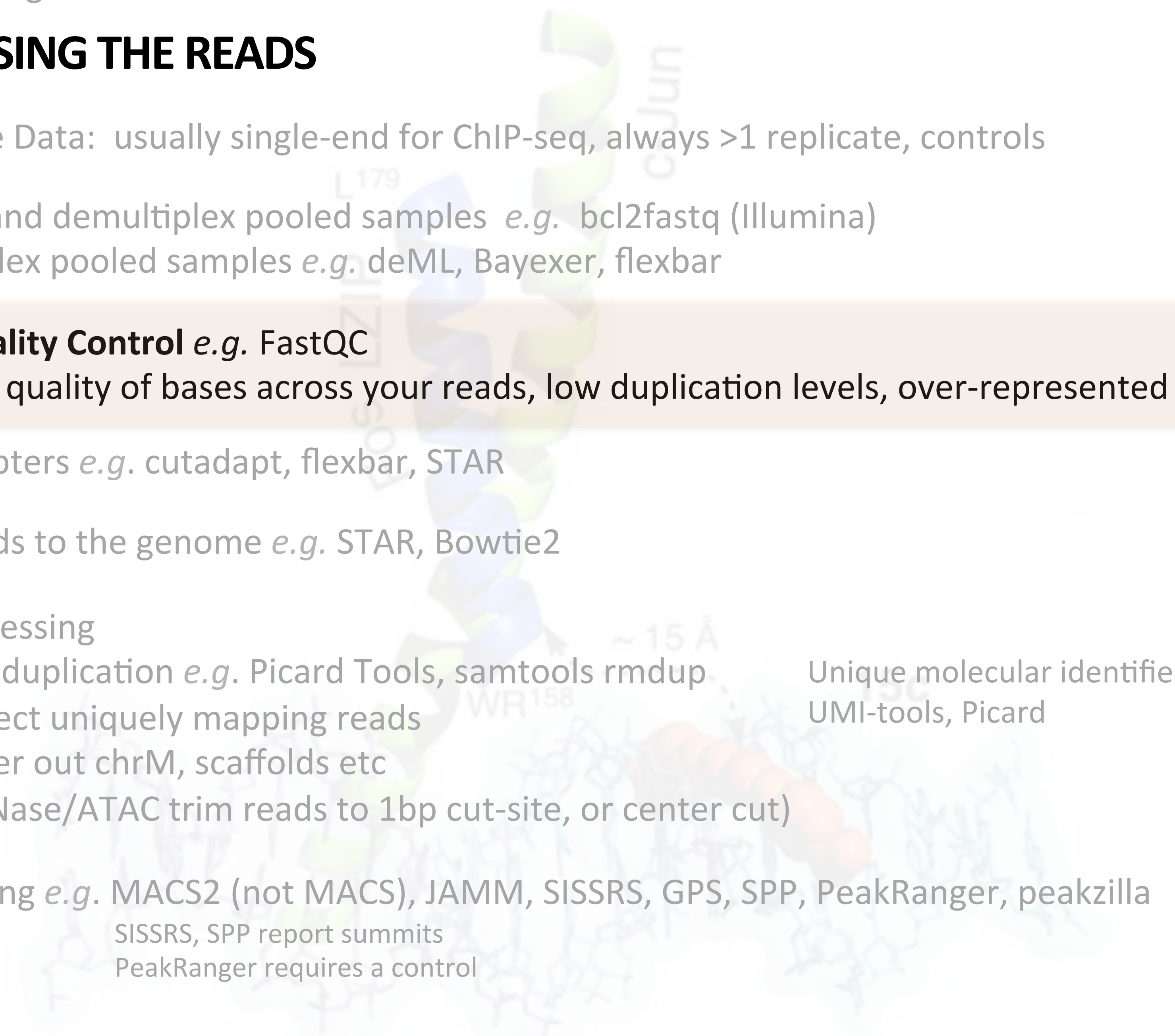
Post processing
- de-duplication *e.g*. Picard Tools, samtools rmdup
- select uniquely mapping reads
- filter out chrM, scaffolds etc
- (DNase/ATAC trim reads to 1bp cut-site, or center cut)

Unique molecular identifiers (UMIs)
UMI-tools, Picard

Peak calling *e.g*. MACS2 (not MACS), JAMM, SISSRS, GPS, SPP, PeakRanger, peakzilla
SISSRS, SPP report summits
PeakRanger requires a control

Filter peaks against the ENCODE blacklist regions

PROCESSING THE READS



Sequence Data: usually single-end for ChIP-seq, always >1 replicate, controls

Convert and demultiplex pooled samples *e.g.* bcl2fastq (Illumina)
Demultiplex pooled samples *e.g.* deML, Bayexer, flexbar

Read Quality Control *e.g.* FastQC
high quality of bases across your reads, low duplication levels, over-represented seqs.

Trim adapters *e.g*. cutadapt, flexbar, STAR

Align reads to the genome *e.g.* STAR, Bowtie2

Post processing
- de-duplication *e.g*. Picard Tools, samtools rmdup
- select uniquely mapping reads
- filter out chrM, scaffolds etc
- (DNase/ATAC trim reads to 1bp cut-site, or center cut)

Unique molecular identifiers (UMIs)
UMI-tools, Picard

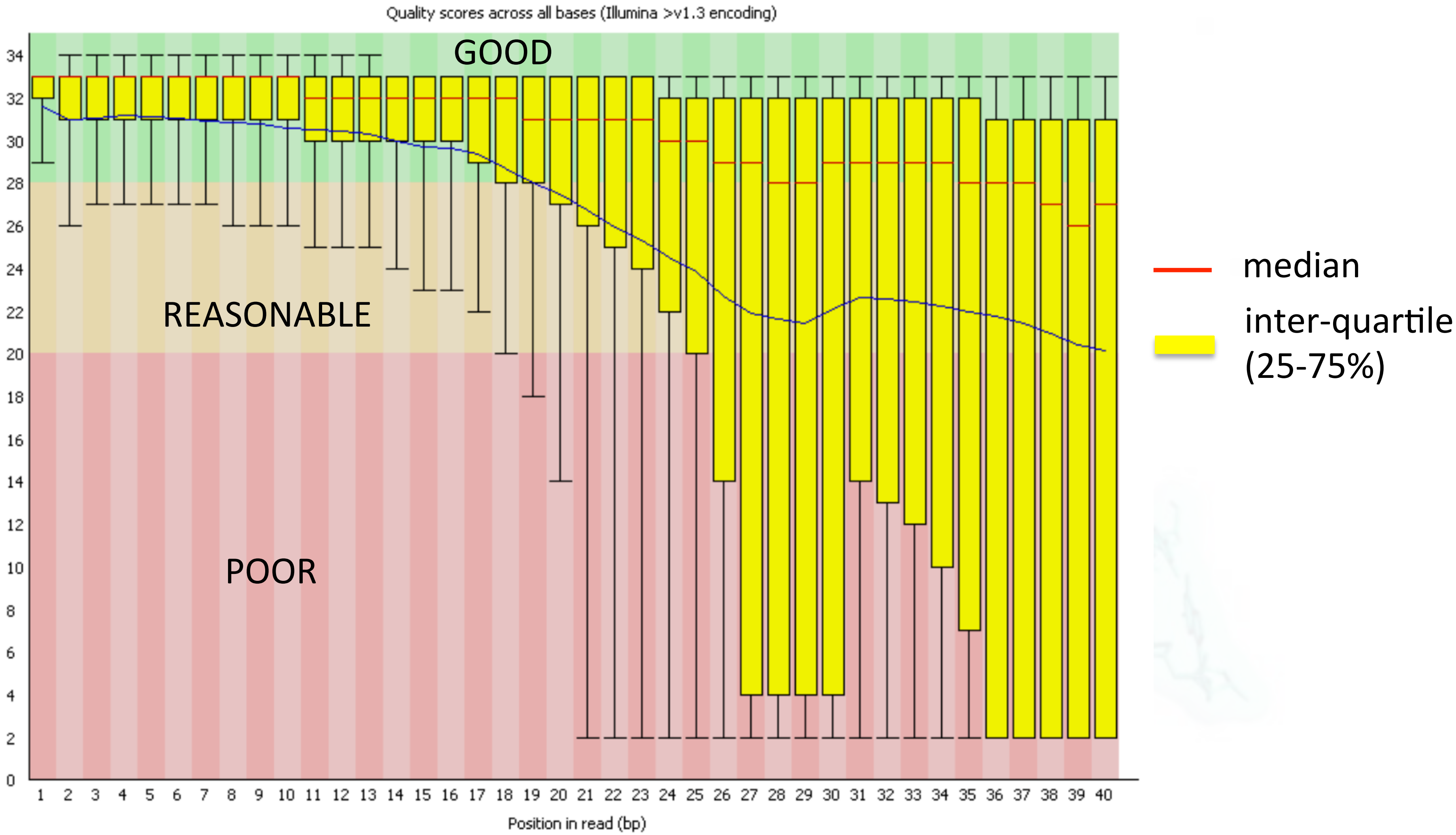
Peak calling *e.g*. MACS2 (not MACS), JAMM, SISSRS, GPS, SPP, PeakRanger, peakzilla

SISSRS, SPP report summits
PeakRanger requires a control

Filter peaks against the ENCODE blacklist regions

Quality control *e.g.* FastQC

high quality of bases across your reads, low duplication levels, over-represented seqs.

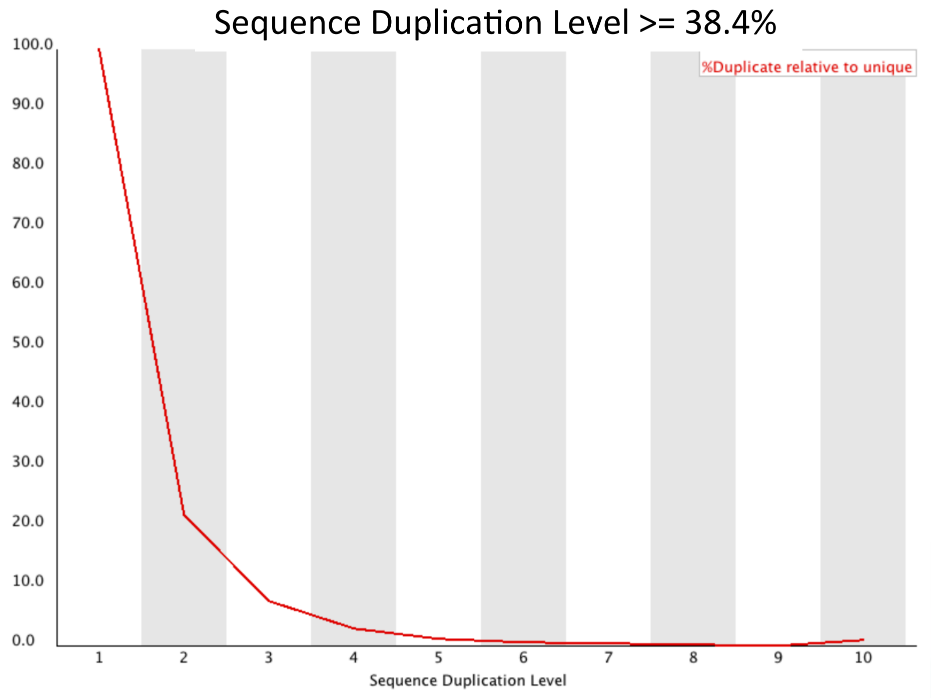


3/4 Peak calling

Quality control *e.g.* FastQC

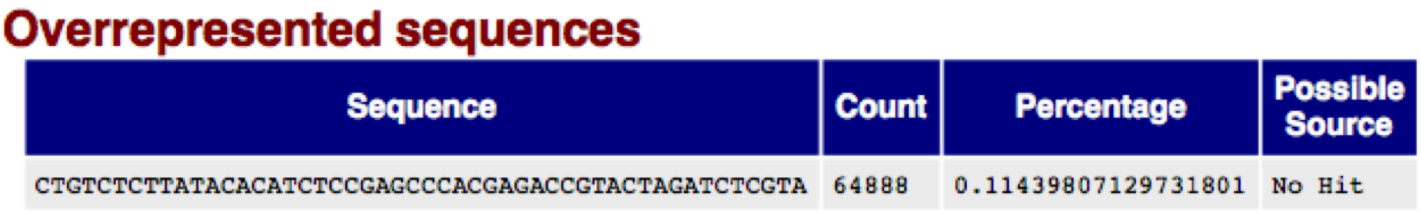
high quality of bases across your reads, **low duplication levels**, **over-represented seqs**.

Sequence Duplication Levels



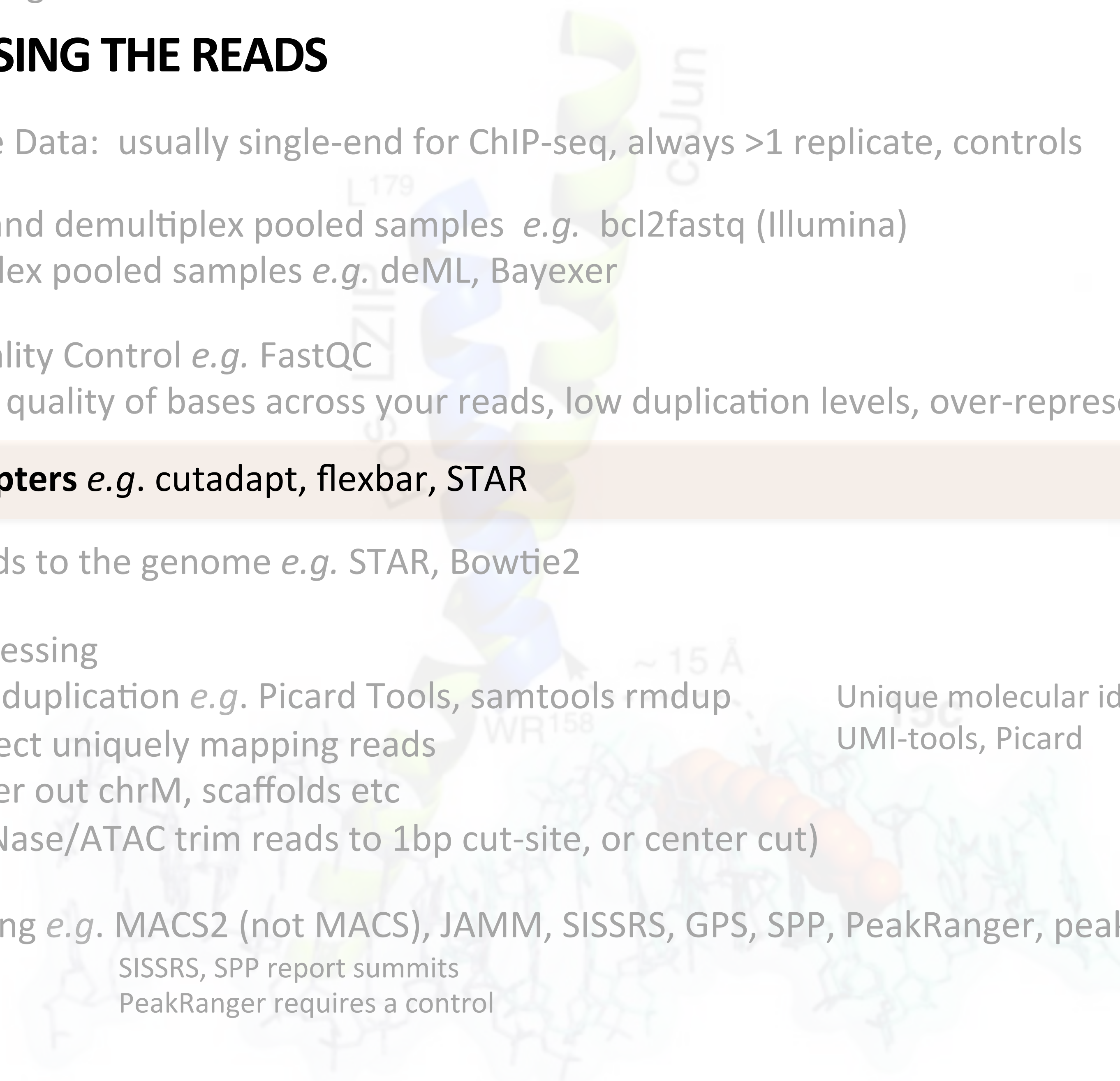
get a warning at 20%

Overrepresented sequences



Sequence	Count	Percentage	Possible Source
CTGTCTCTTATACACATCTCCGAGCCCACGAGACCGTACTAGATCTCGTA	64888	0.11439807129731801	No Hit

PROCESSING THE READS



Sequence Data: usually single-end for ChIP-seq, always >1 replicate, controls

Convert and demultiplex pooled samples *e.g.* bcl2fastq (Illumina)
Demultiplex pooled samples *e.g.* deML, Bayexer

Read Quality Control *e.g.* FastQC
high quality of bases across your reads, low duplication levels, over-represented seqs.

Trim adapters *e.g*. cutadapt, flexbar, STAR

Align reads to the genome *e.g.* STAR, Bowtie2

Post processing
- de-duplication *e.g*. Picard Tools, samtools rmdup
- select uniquely mapping reads
- filter out chrM, scaffolds etc
- (DNase/ATAC trim reads to 1bp cut-site, or center cut)

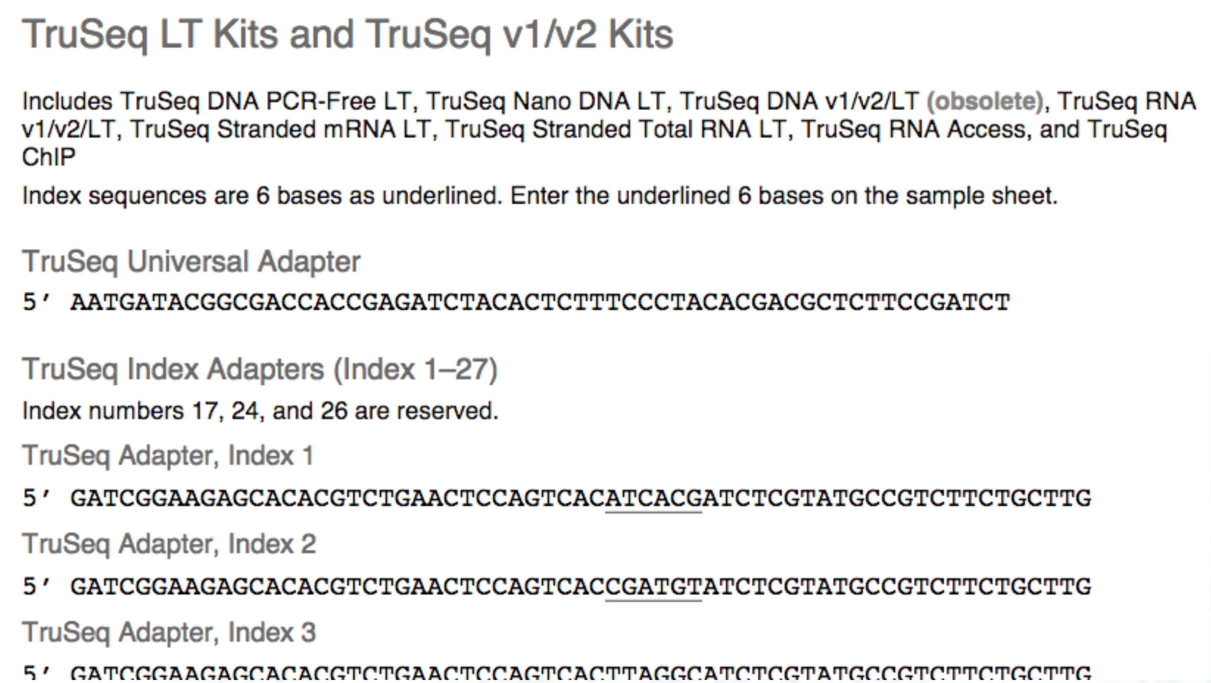
Unique molecular identifiers (UMIs)
UMI-tools, Picard

Peak calling *e.g*. MACS2 (not MACS), JAMM, SISSRS, GPS, SPP, PeakRanger, peakzilla
SISSRS, SPP report summits
PeakRanger requires a control

Filter peaks against the ENCODE blacklist regions

Trim adapters *e.g*. cutadapt, flexbar

Yours: you know what you used
Public data: hopefully they reported adapters
Unknown: from FastQC – search for worst offenders
“Illumina Adapter Sequences Document”



TruSeq LT Kits and TruSeq v1/v2 Kits

Includes TruSeq DNA PCR-Free LT, TruSeq Nano DNA LT, TruSeq DNA v1/v2/LT (obsolete), TruSeq RNA v1/v2/LT, TruSeq Stranded mRNA LT, TruSeq Stranded Total RNA LT, TruSeq RNA Access, and TruSeq ChIP

Index sequences are 6 bases as underlined. Enter the underlined 6 bases on the sample sheet.

TruSeq Universal Adapter

5′ AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

TruSeq Index Adapters (Index 1–27)

Index numbers 17, 24, and 26 are reserved.

TruSeq Adapter, Index 1

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG

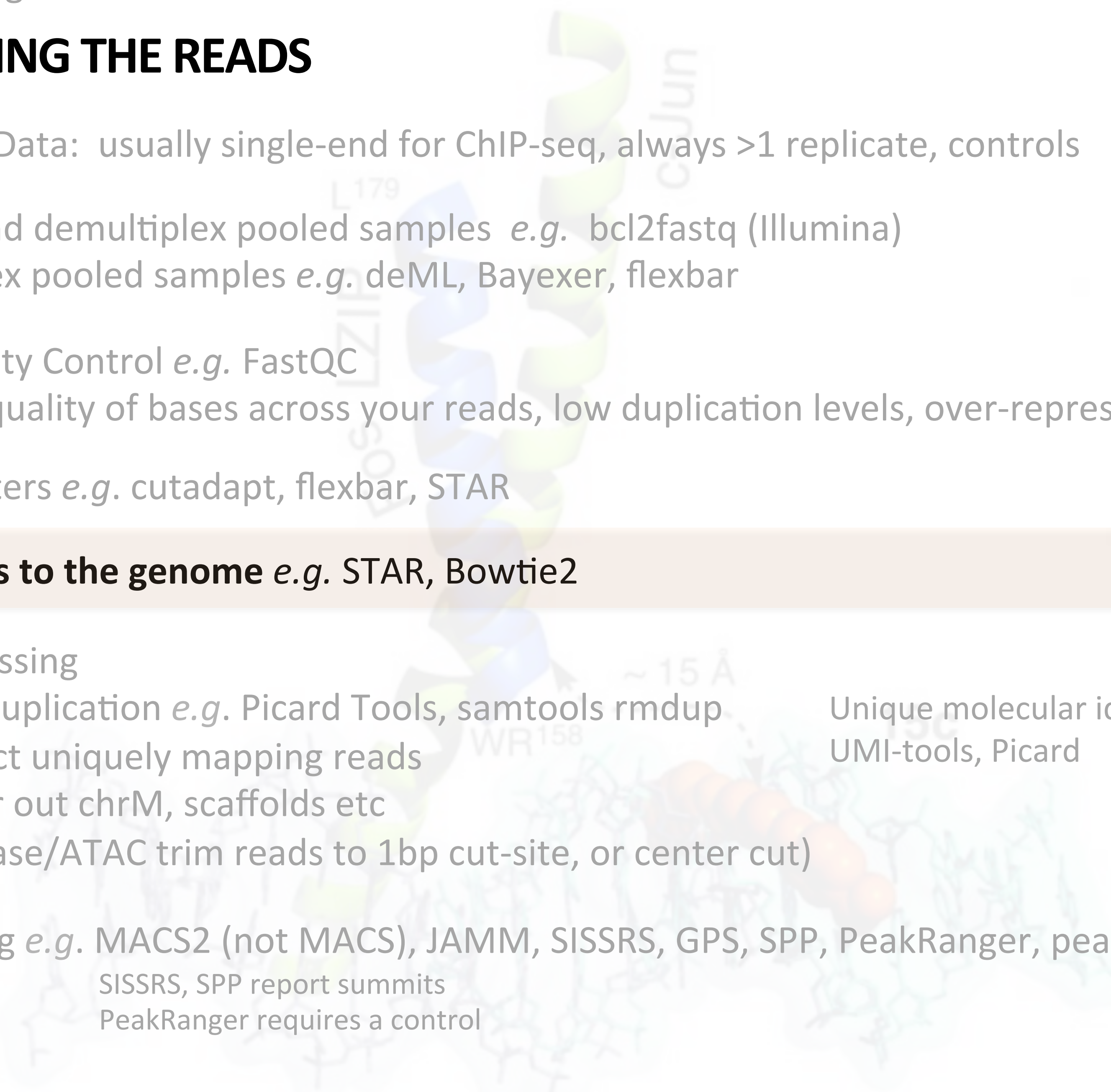
TruSeq Adapter, Index 2

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTG

TruSeq Adapter, Index 3

5′ GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG

PROCESSING THE READS



Sequence Data: usually single-end for ChIP-seq, always >1 replicate, controls

Convert and demultiplex pooled samples *e.g.* bcl2fastq (Illumina)
Demultiplex pooled samples *e.g.* deML, Bayexer, flexbar

Read Quality Control *e.g.* FastQC
high quality of bases across your reads, low duplication levels, over-represented seqs.

Trim adapters *e.g*. cutadapt, flexbar, STAR

Align reads to the genome *e.g.* STAR, Bowtie2

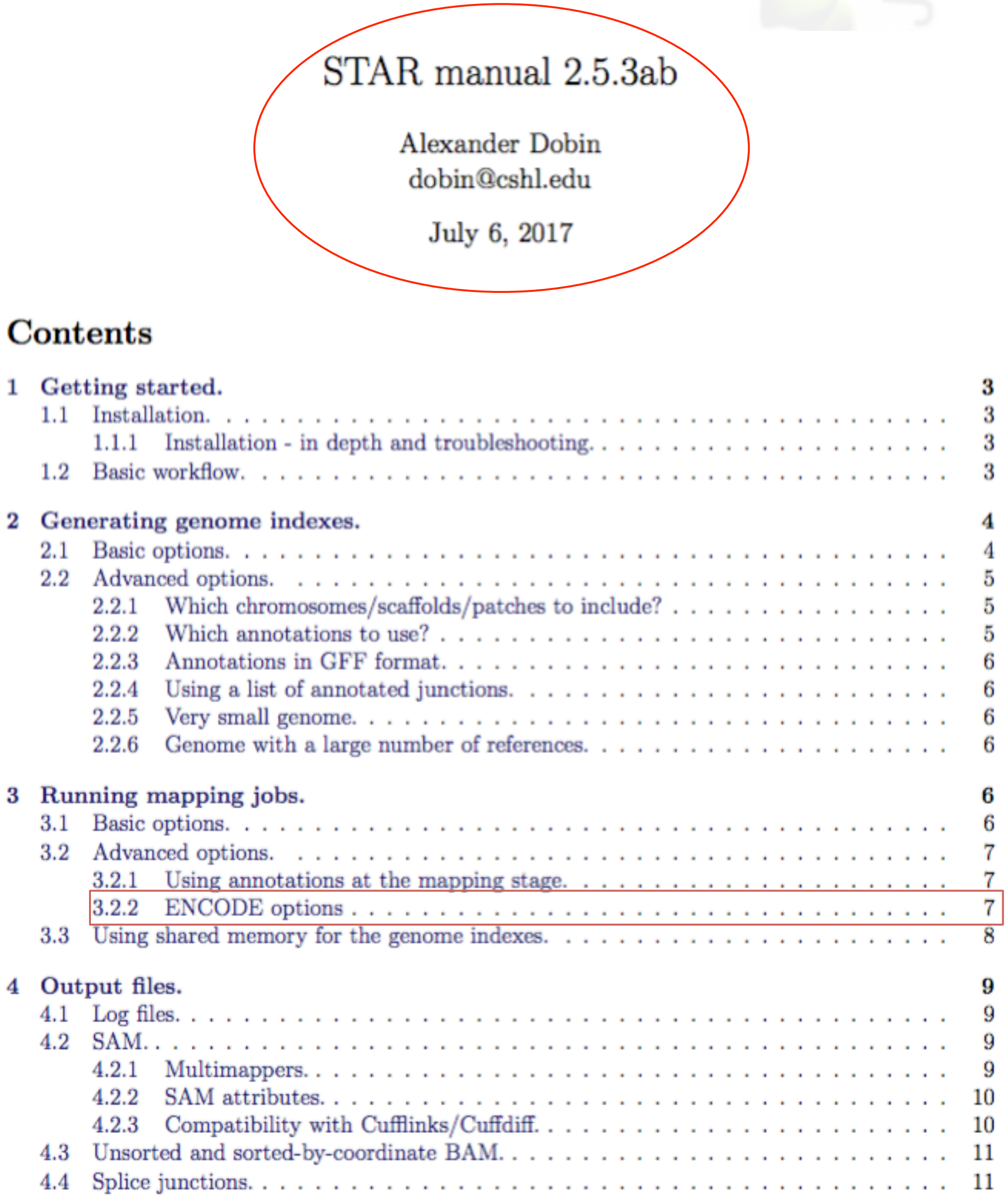
Post processing
- de-duplication *e.g*. Picard Tools, samtools rmdup
- select uniquely mapping reads
- filter out chrM, scaffolds etc
- (DNase/ATAC trim reads to 1bp cut-site, or center cut)

Unique molecular identifiers (UMIs)
UMI-tools, Picard

Peak calling *e.g*. MACS2 (not MACS), JAMM, SISSRS, GPS, SPP, PeakRanger, peakzilla
SISSRS, SPP report summits
PeakRanger requires a control

Filter peaks against the ENCODE blacklist regions

Align reads to the genome *e.g.* STAR, Bowtie2



STAR manual 2.5.3ab

Alexander Dobin
dobin@cshl.edu

July 6, 2017

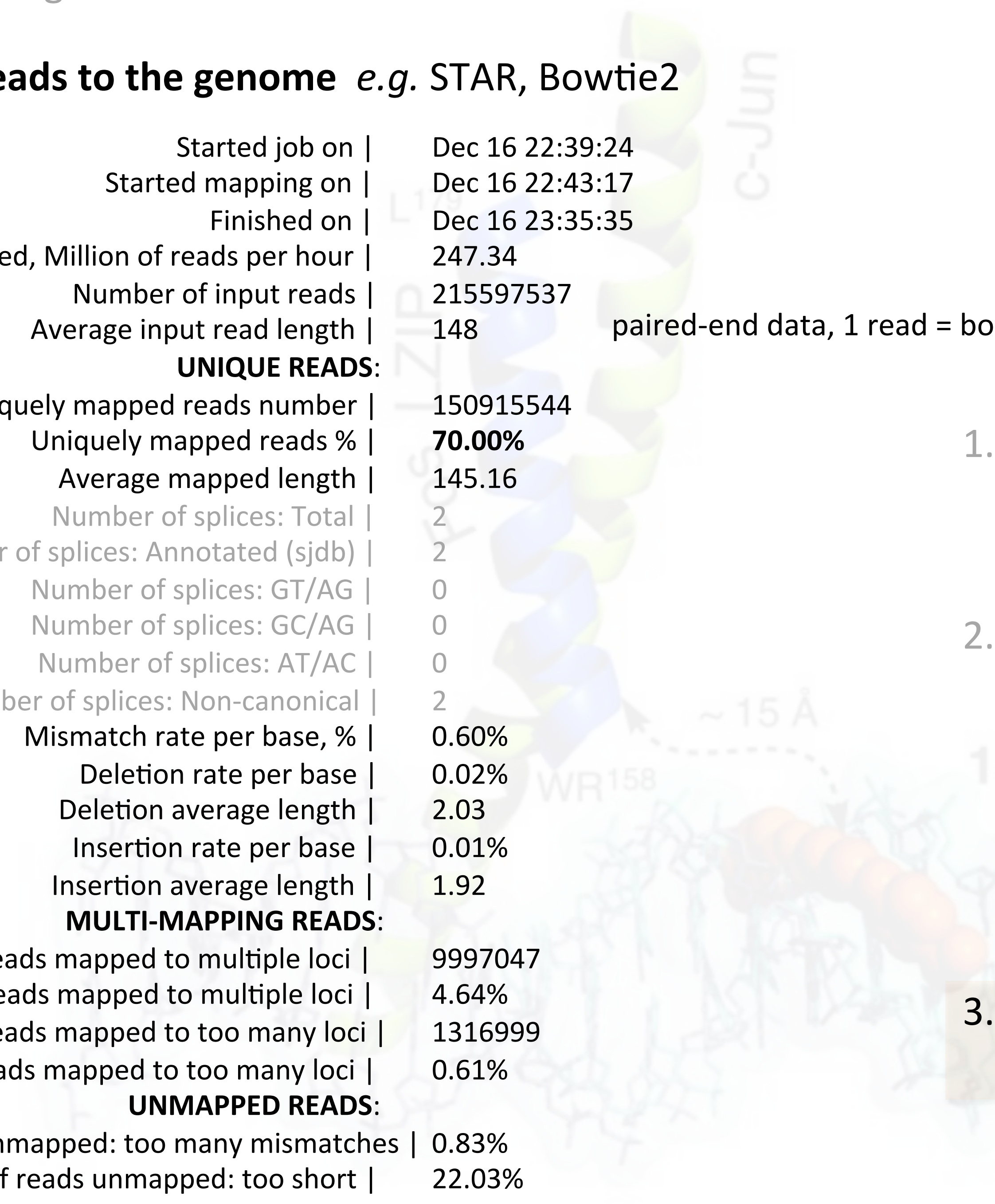
Contents

1. Generate genome index
   - reference genome
   - gtf annotation file

2. Map reads to genome
   - genome index
   - fastq/fasta files
   - for ChIP/DNase/ATAC-seq suppress splice junctions
   - if paired end, maybe limit gap between mates

3. Check statistic file
   *_Log.final.out

Align reads to the genome *e.g.* STAR, Bowtie2



```
                                 Started job on |      Dec 16 22:39:24
                             Started mapping on |      Dec 16 22:43:17
                                    Finished on |      Dec 16 23:35:35
       Mapping speed, Million of reads per hour |      247.34
                          Number of input reads |      215597537
                      Average input read length |      148
                                    UNIQUE READS:
                   Uniquely mapped reads number |      150915544
                        Uniquely mapped reads % |      70.00%
                          Average mapped length |      145.16
                       Number of splices: Total |      2
            Number of splices: Annotated (sjdb) |      2
                       Number of splices: GT/AG |      0
                       Number of splices: GC/AG |      0
                       Number of splices: AT/AC |      0
               Number of splices: Non-canonical |      2
                      Mismatch rate per base, % |      0.60%
                         Deletion rate per base |      0.02%
                        Deletion average length |      2.03
                        Insertion rate per base |      0.01%
                       Insertion average length |      1.92
                             MULTI-MAPPING READS:
        Number of reads mapped to multiple loci |      9997047
             % of reads mapped to multiple loci |      4.64%
        Number of reads mapped to too many loci |      1316999
             % of reads mapped to too many loci |      0.61%
                                  UNMAPPED READS:
       % of reads unmapped: too many mismatches |      0.83%
                 % of reads unmapped: too short |      22.03%
                     % of reads unmapped: other |      1.89%
```

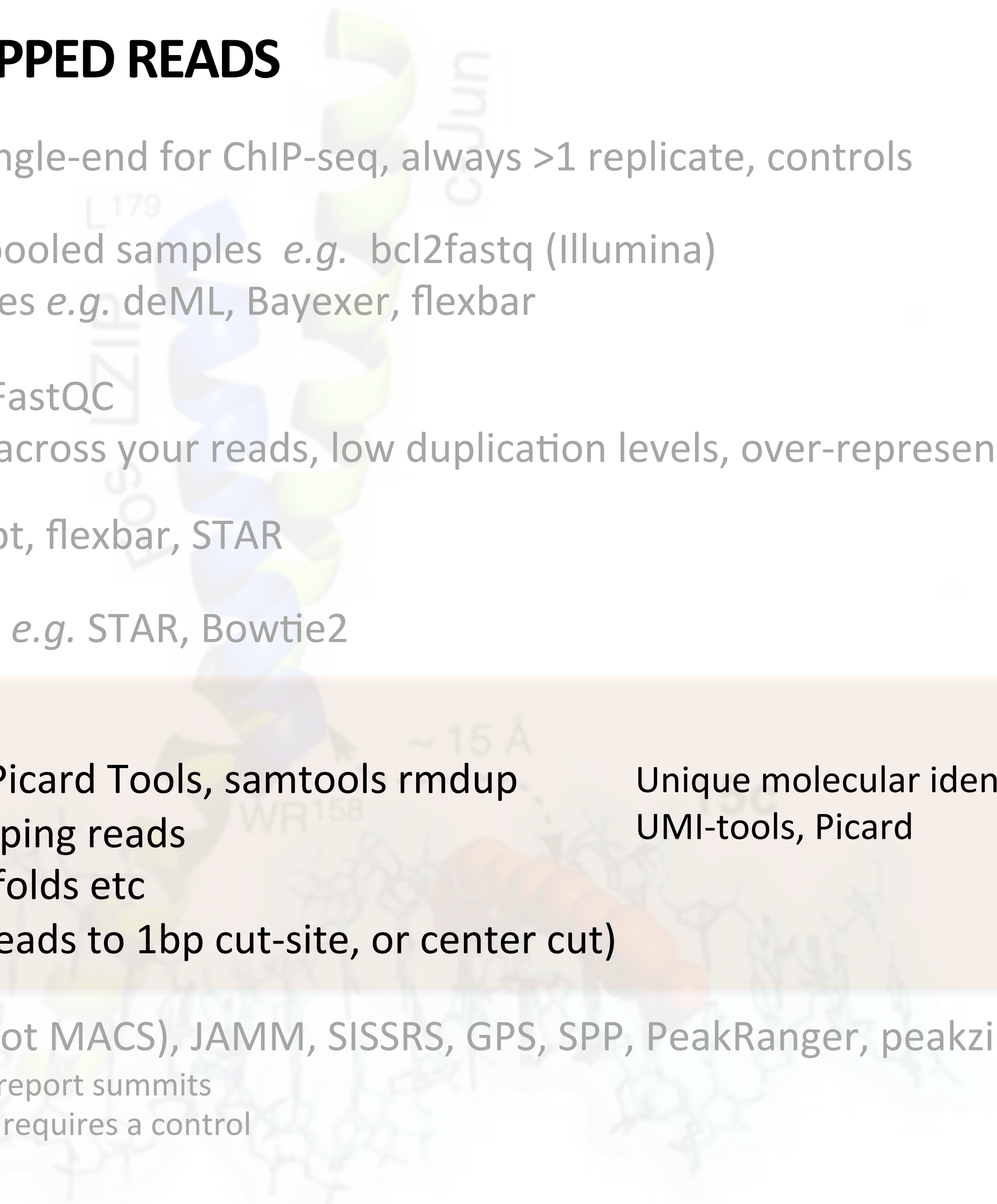
paired-end data, 1 read = both mates

1. Generate genome index
   - reference genome
   - gtf annotation file

2. Map reads to genome
   - genome index
   - fastq/fasta files
   - for ChIP/DNase/ATAC-seq suppress splice junctions
   - if paired end, maybe limit gap between mates

3. Check statistic file
   *_Log.final.out

PROCESSING THE MAPPED READS



Sequence Data: usually single-end for ChIP-seq, always >1 replicate, controls

Convert and demultiplex pooled samples *e.g.* bcl2fastq (Illumina)
Demultiplex pooled samples *e.g.* deML, Bayexer, flexbar

Read Quality Control *e.g.* FastQC
high quality of bases across your reads, low duplication levels, over-represented seqs.

Trim adapters *e.g*. cutadapt, flexbar, STAR

Align reads to the genome *e.g.* STAR, Bowtie2

Post processing

- de-duplication *e.g*. Picard Tools, samtools rmdup
- select uniquely mapping reads
- filter out chrM, scaffolds etc
- (DNase/ATAC trim reads to 1bp cut-site, or center cut)

Unique molecular identifiers (UMIs)
UMI-tools, Picard

Peak calling *e.g*. MACS2 (not MACS), JAMM, SISSRS, GPS, SPP, PeakRanger, peakzilla
SISSRS, SPP report summits
PeakRanger requires a control

Filter peaks against the ENCODE blacklist regions

Post processing mapped data

- de-duplication *e.g*. Picard Tools, samtools rmdup
- select uniquely mapping reads
- filter out chrM, scaffolds etc
- (paired-end concordant reads)
- (DNase/ATAC trim reads to 1bp cut-site, or center cut)

Uniquely mapped:

STAR:

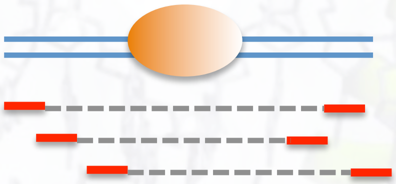
keep uniquely mapped reads

```
awk '{if( $0 ~ /^@/ || $0 ~ /NH:i:1\t/) print $0 }' $mapped > $outfile
```

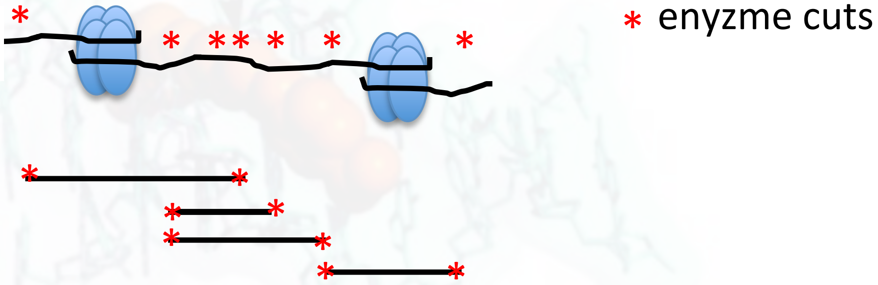
Bowtie2:

keep things that mapped, then remove multi-mappers, also limit number of mis-matches

Keep reads or trim reads:

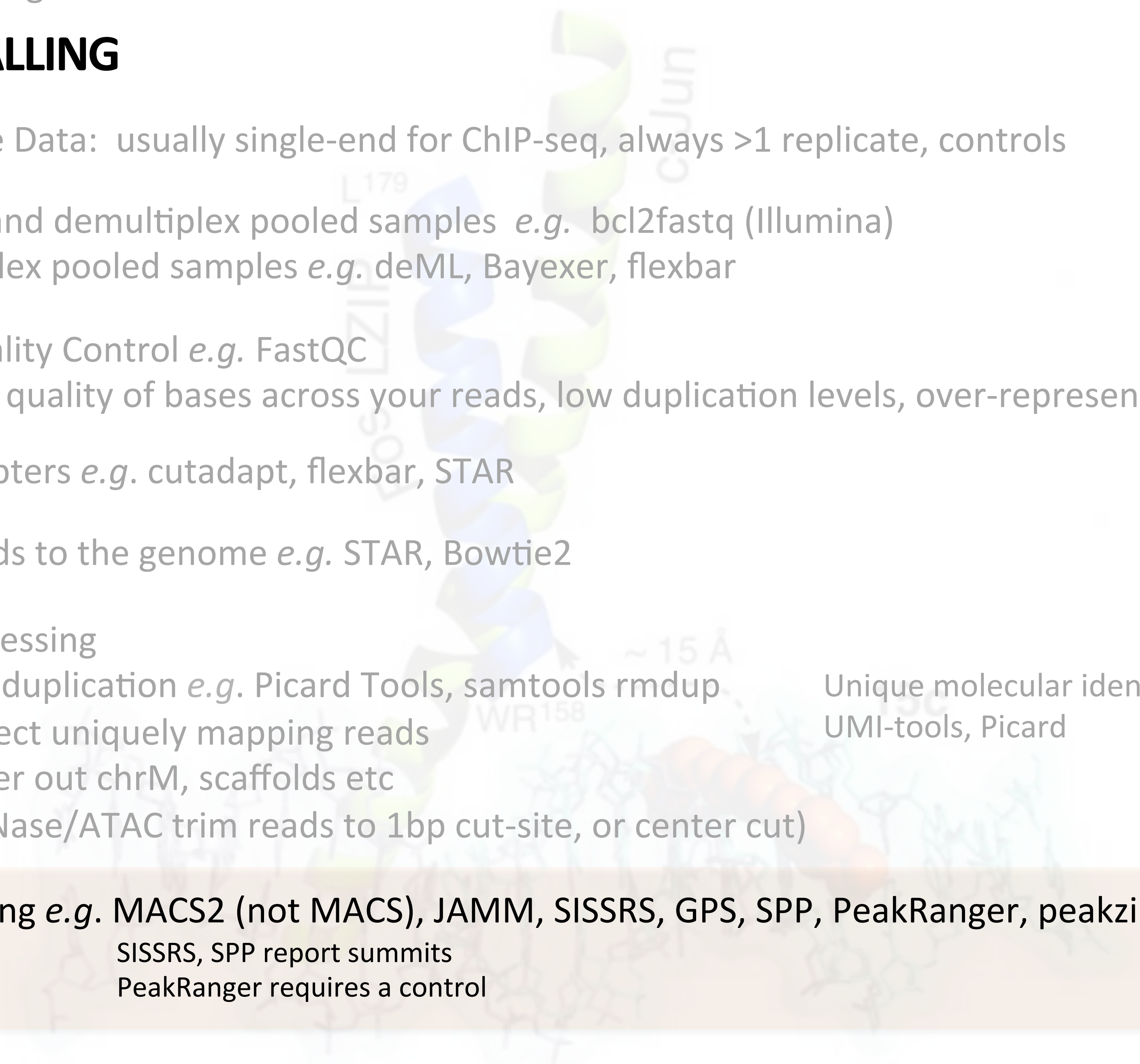


keep reads for ChIP-seq
estimate fragments



ATAC-seq or DNase-seq
know fragments, but keep cut-sites

PEAK CALLING



Sequence Data: usually single-end for ChIP-seq, always >1 replicate, controls

Convert and demultiplex pooled samples *e.g.* bcl2fastq (Illumina)
Demultiplex pooled samples *e.g.* deML, Bayexer, flexbar

Read Quality Control *e.g.* FastQC
    high quality of bases across your reads, low duplication levels, over-represented seqs.

Trim adapters *e.g*. cutadapt, flexbar, STAR

Align reads to the genome *e.g.* STAR, Bowtie2

Post processing
- de-duplication *e.g*. Picard Tools, samtools rmdup
- select uniquely mapping reads
- filter out chrM, scaffolds etc
- (DNase/ATAC trim reads to 1bp cut-site, or center cut)

Unique molecular identifiers (UMIs)
UMI-tools, Picard

Peak calling *e.g*. MACS2 (not MACS), JAMM, SISSRS, GPS, SPP, PeakRanger, peakzilla
    SISSRS, SPP report summits
    PeakRanger requires a control

Filter peaks against the ENCODE blacklist regions

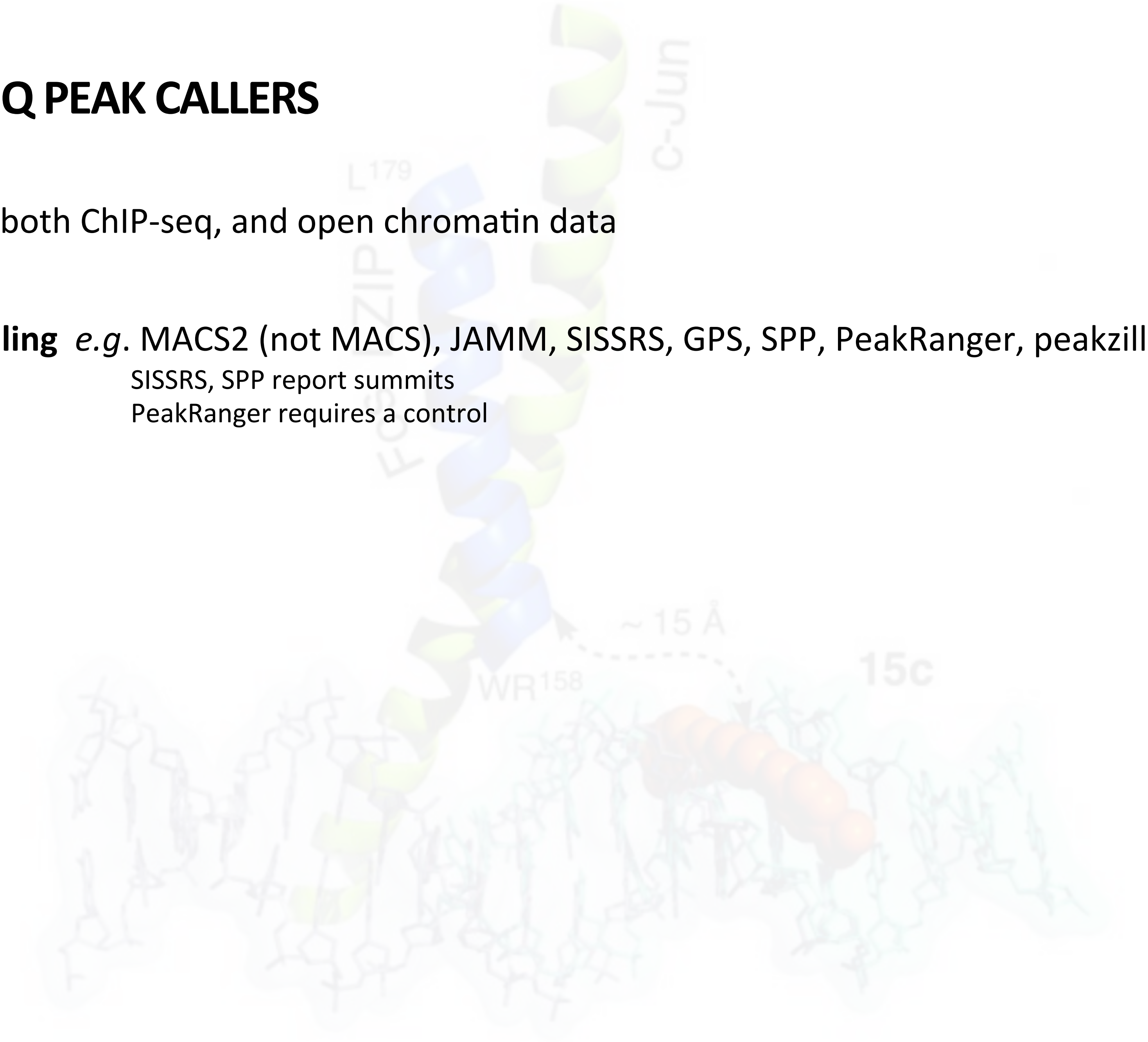
ChIP-SEQ PEAK CALLERS

Good for both ChIP-seq, and open chromatin data

Peak calling *e.g*. MACS2 (not MACS), JAMM, SISSRS, GPS, SPP, PeakRanger, peakzilla

SISSRS, SPP report summits

PeakRanger requires a control



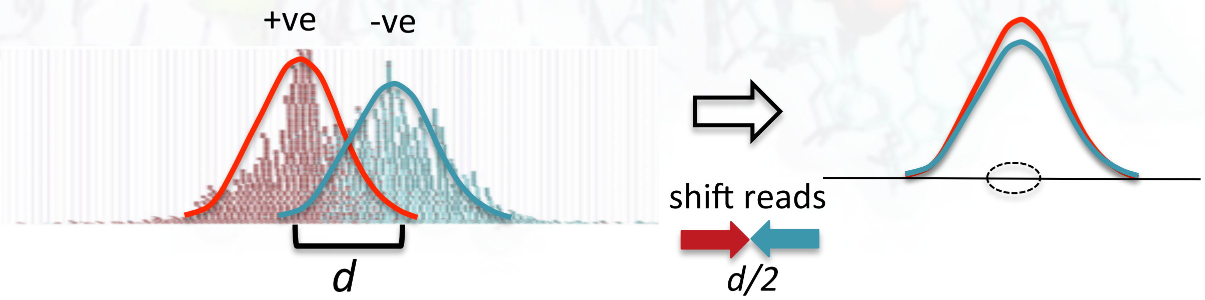
MACS2

- MACS2, not MACS
- One of the classic ChIP-seq peak callers

- Fragment size estimation model:
  - *Bandwidth* = sonication size
  - *mfold, default: 5,50* = range of required fold-enrichment over background

Can choose a fixed fragment size

  - Identify regions of size *2*bandwidth* with tags more than *mfold* enriched relative to a random distribution
  - Sample 1,000 of these high-quality read enriched regions → separate +ve and –ve strands
  - Shift all tags by *d/2* toward the 3' ends

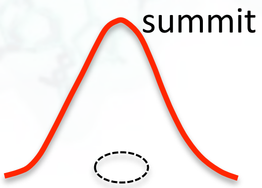


Zhang et al, Genome Biol 2008

MACS2: FURTHER STEPS

- Fragment size model now used to evaluate the entire dataset
- How do genomic regions compare to the assumption of random distribution?
  - Slide *2*fragmentsize* windows across the genome, assess the distribution of tags against a Poisson distribution
  - significant tag enrichment = Poisson *p*-value based on λ_{BG} , default 10e-5
    - Poisson advantage is that λ captures both mean and variance of the distribution

- Overlapping enriched regions are merged
- Location with highest fragment pileup is predicted as the protein binding location



MACS2: ESTIMATING BACKGROUND

- Background estimate is the input control sample, or the ChIP-Seq sample itself
- Poisson: $$P(K=k)=\frac{\lambda^k e^{-\lambda}}{k!}$$

  k is the number of times an event occurs in an interval (want prob. of k)
  λ is the average number of events in the interval (known)
  P the probability of k events happening in the interval

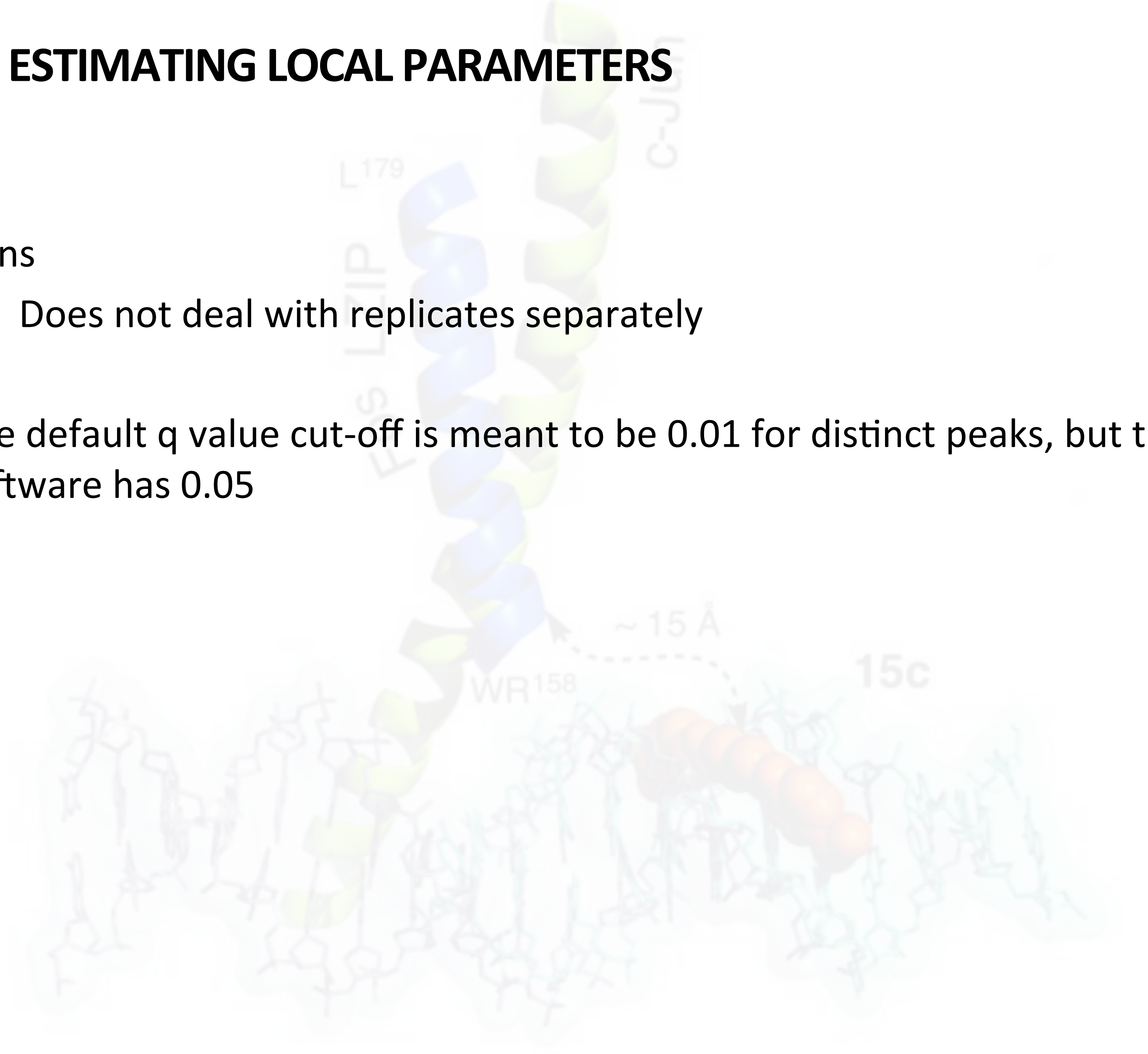
- The background used for a given peak comes from a max of several sources of λ:

$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k})$$

  - λ_{1k}, λ_{5k} and λ_{10k} are the λ estimated from the 1 kb, 5 kb or 10 kb window centered at the peak location (if no input control, omit λ_{1k})
  - λ_{BG} is from the whole genome

- Ranks peaks by ratio of (peak tag count) / λ_{local}
- Calculate FDR via sample swap *i.e.* control data/peak data

MACS2: ESTIMATING LOCAL PARAMETERS

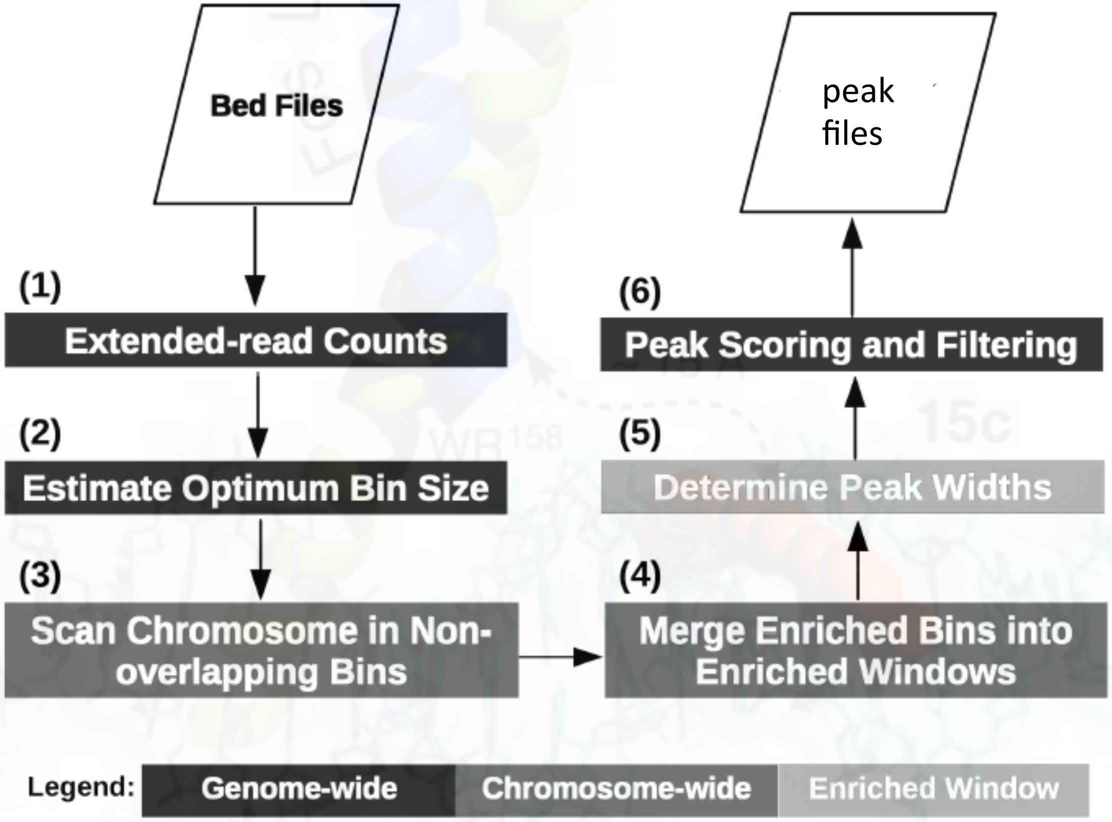


- Cons
  - Does not deal with replicates separately

- The default q value cut-off is meant to be 0.01 for distinct peaks, but the software has 0.05

JAMM

- Joint analysis of replicates via multivariate mixtures
- High resolution detection of peak edges



Bed Files

(1) Extended-read Counts

(2) Estimate Optimum Bin Size

(3) Scan Chromosome in Non-overlapping Bins

(4) Merge Enriched Bins into Enriched Windows

(5) Determine Peak Widths

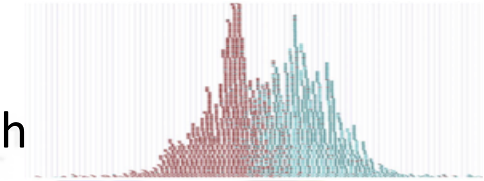
(6) Peak Scoring and Filtering

peak files

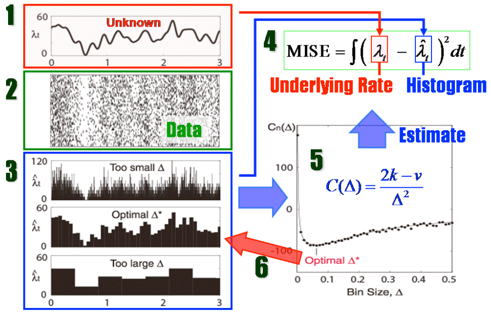
Legend: Genome-wide | Chromosome-wide | Enriched Window

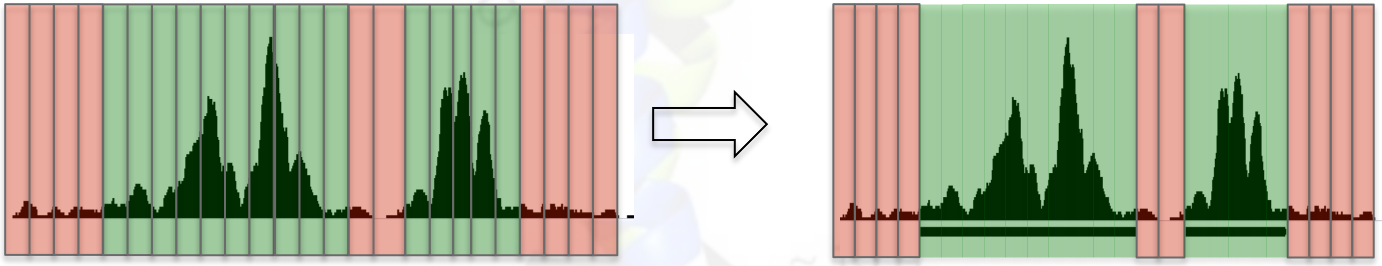
JAMM

- Estimate fragment length

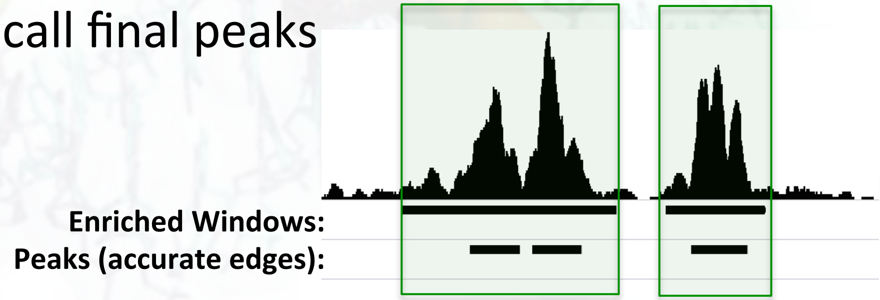


- Break genome into bins to assess enriched or background
  - Goldilocks bins: don't want bins too small, nor too big



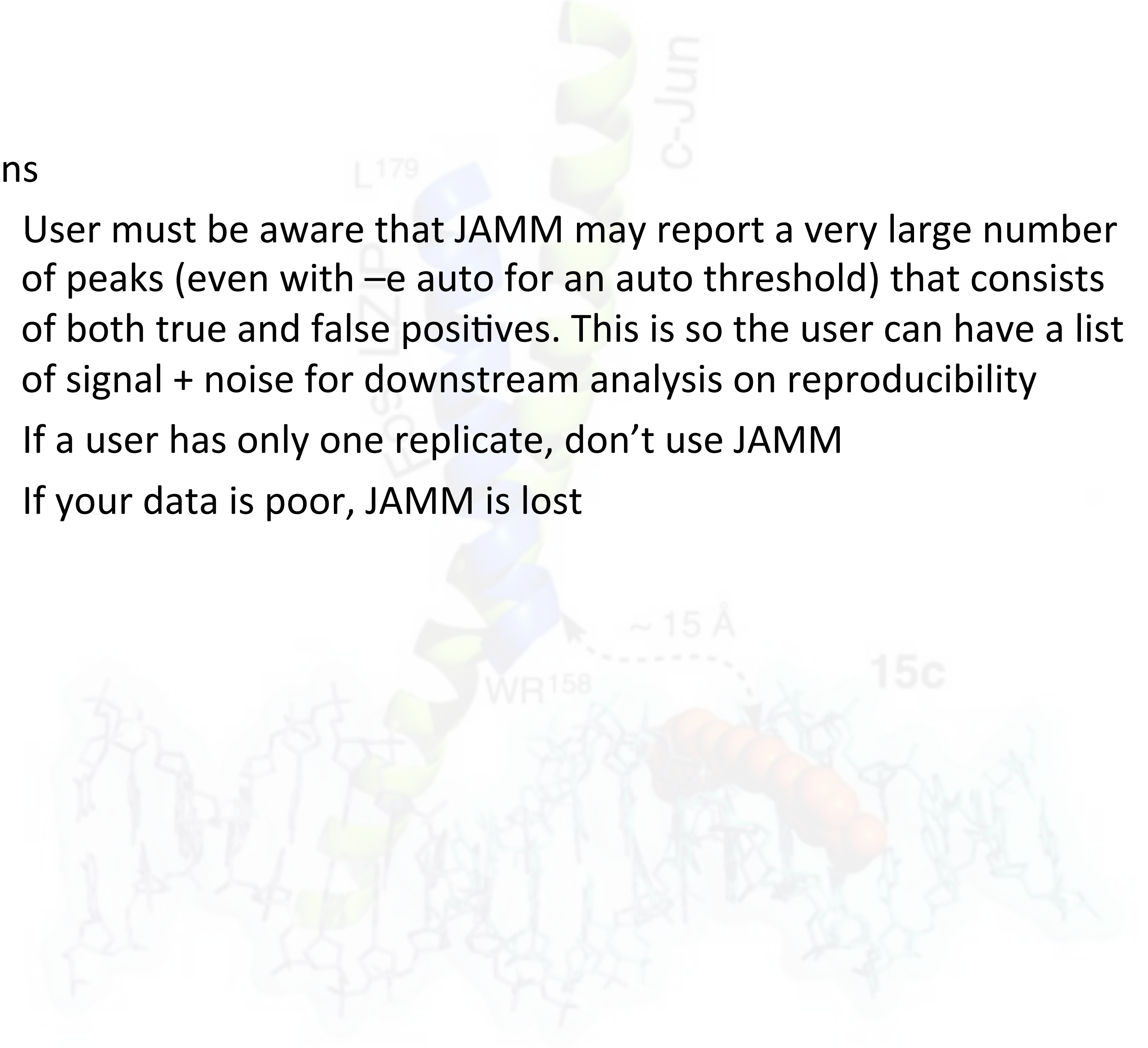


- Within a window use Multivariate Gaussian Mixture Model Clustering to break window into read enriched vs background
- Replicates are then used to call final peaks



JAMM

- Cons
  - User must be aware that JAMM may report a very large number of peaks (even with –e auto for an auto threshold) that consists of both true and false positives. This is so the user can have a list of signal + noise for downstream analysis on reproducibility
  - If a user has only one replicate, don’t use JAMM
  - If your data is poor, JAMM is lost



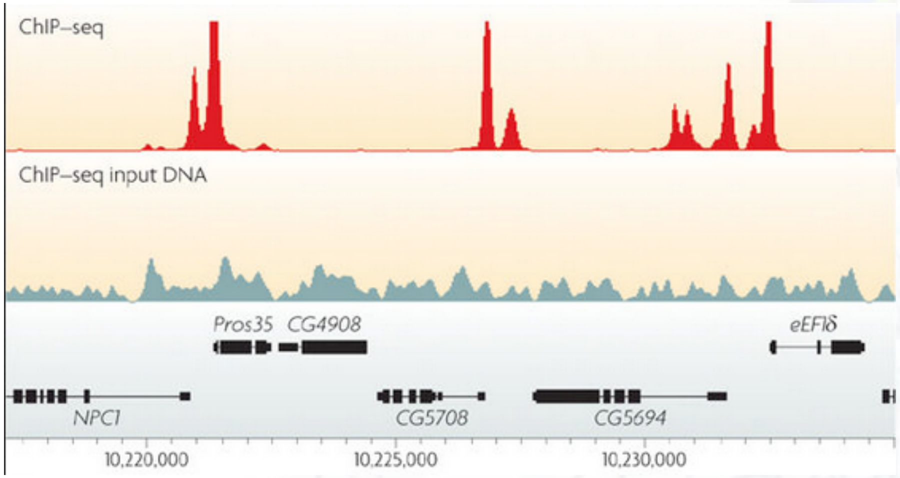
PEAKS: DISTINCT REGIONS AND BROAD REGIONS

Distinct peaks: generally < 800bp

Non-nucleosome proteins
e.g. TFs, remodelers, enzymes
Some histone modifications
e.g. H3K4me3

Broad peaks: 100's of kb

Some histone modifications
e.g. H3K27me3, H3K9me3
H3K9me1, H3K36me3



Sims et al Nature Reviews Genetics 15, 121–132 (2014)

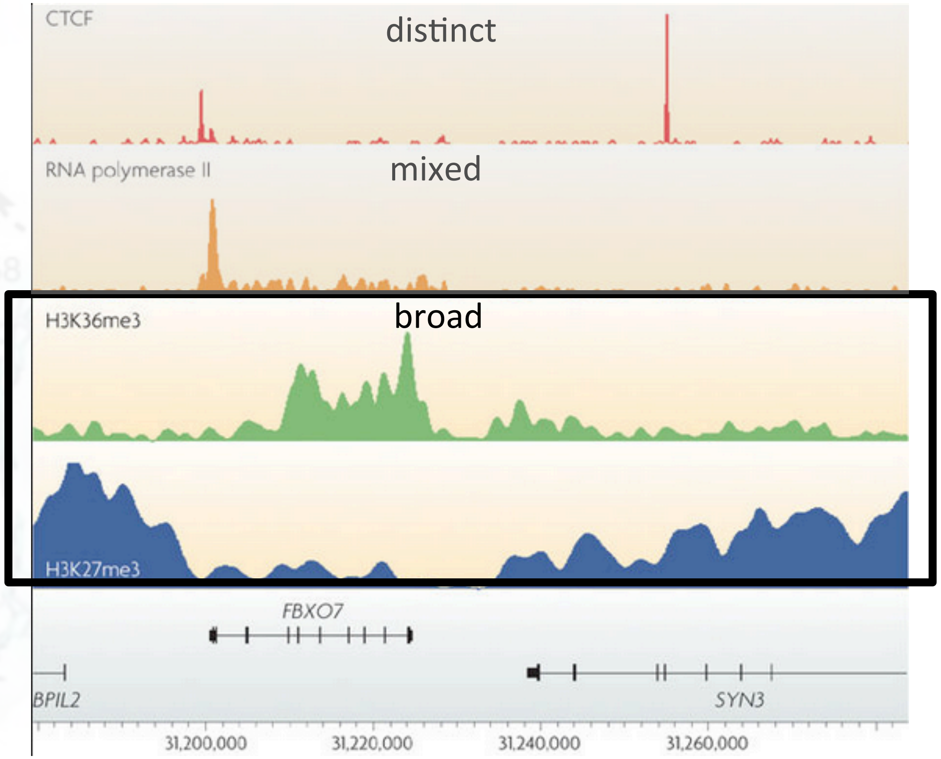
DISTINCT REGIONS AND BROAD REGIONS

Distinct peaks: generally < 800bp

Non-nucleosome proteins
e.g. TFs, remodelers, enzymes
Some histone modifications
e.g. H3K4me3

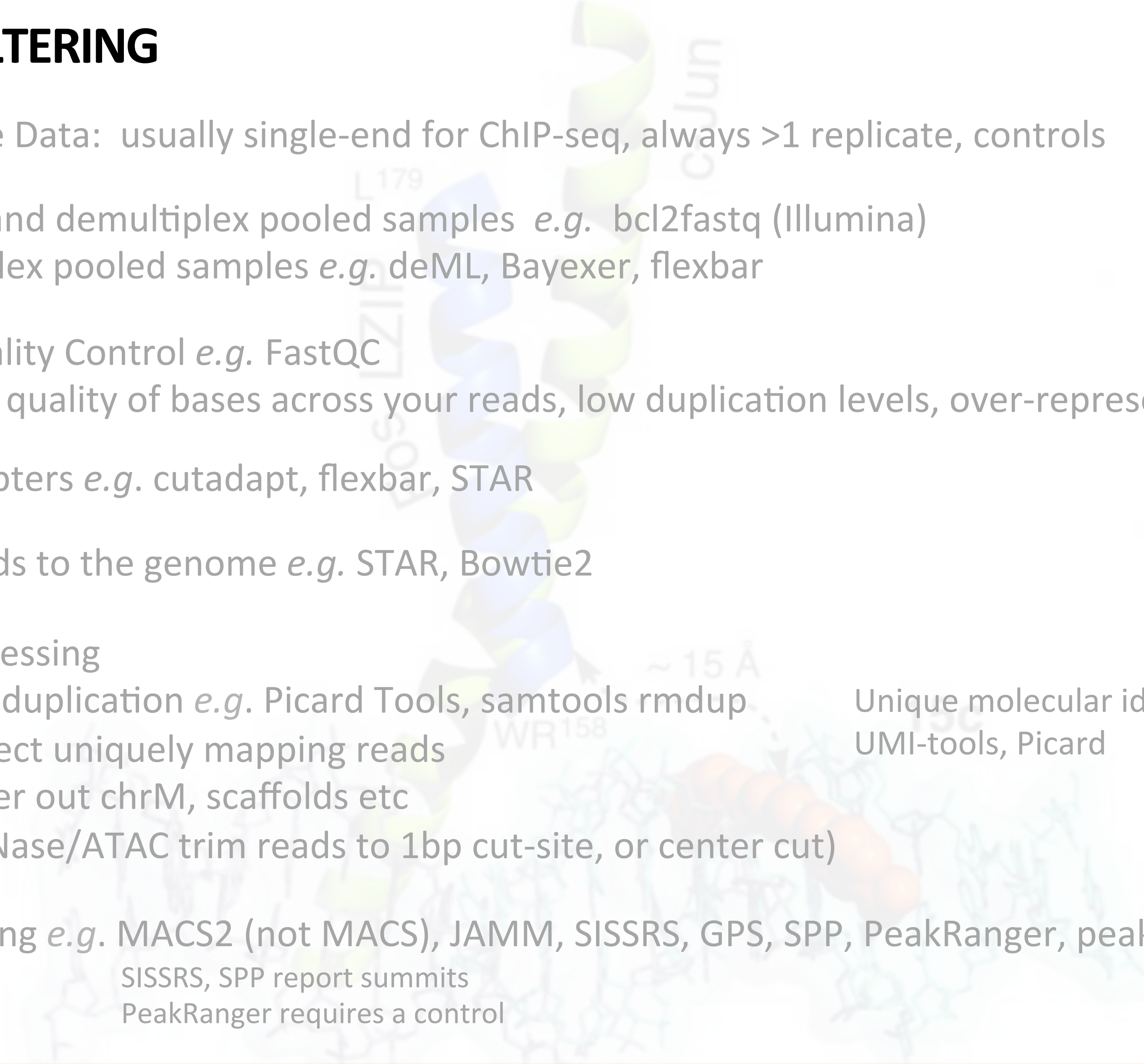
Broad peaks: 100’s of kb

Some histone modifications
e.g. H3K27me3, H3K9me3
H3K9me1, H3K36me3



Sims et al Nature Reviews Genetics 15, 121–132 (2014)

PEAK FILTERING



Sequence Data: usually single-end for ChIP-seq, always >1 replicate, controls

Convert and demultiplex pooled samples *e.g.* bcl2fastq (Illumina)
Demultiplex pooled samples *e.g.* deML, Bayexer, flexbar

Read Quality Control *e.g.* FastQC
high quality of bases across your reads, low duplication levels, over-represented seqs.

Trim adapters *e.g*. cutadapt, flexbar, STAR

Align reads to the genome *e.g.* STAR, Bowtie2

Post processing
- de-duplication *e.g*. Picard Tools, samtools rmdup
- select uniquely mapping reads
- filter out chrM, scaffolds etc
- (DNase/ATAC trim reads to 1bp cut-site, or center cut)

Unique molecular identifiers (UMIs)
UMI-tools, Picard

Peak calling *e.g*. MACS2 (not MACS), JAMM, SISSRS, GPS, SPP, PeakRanger, peakzilla
SISSRS, SPP report summits
PeakRanger requires a control

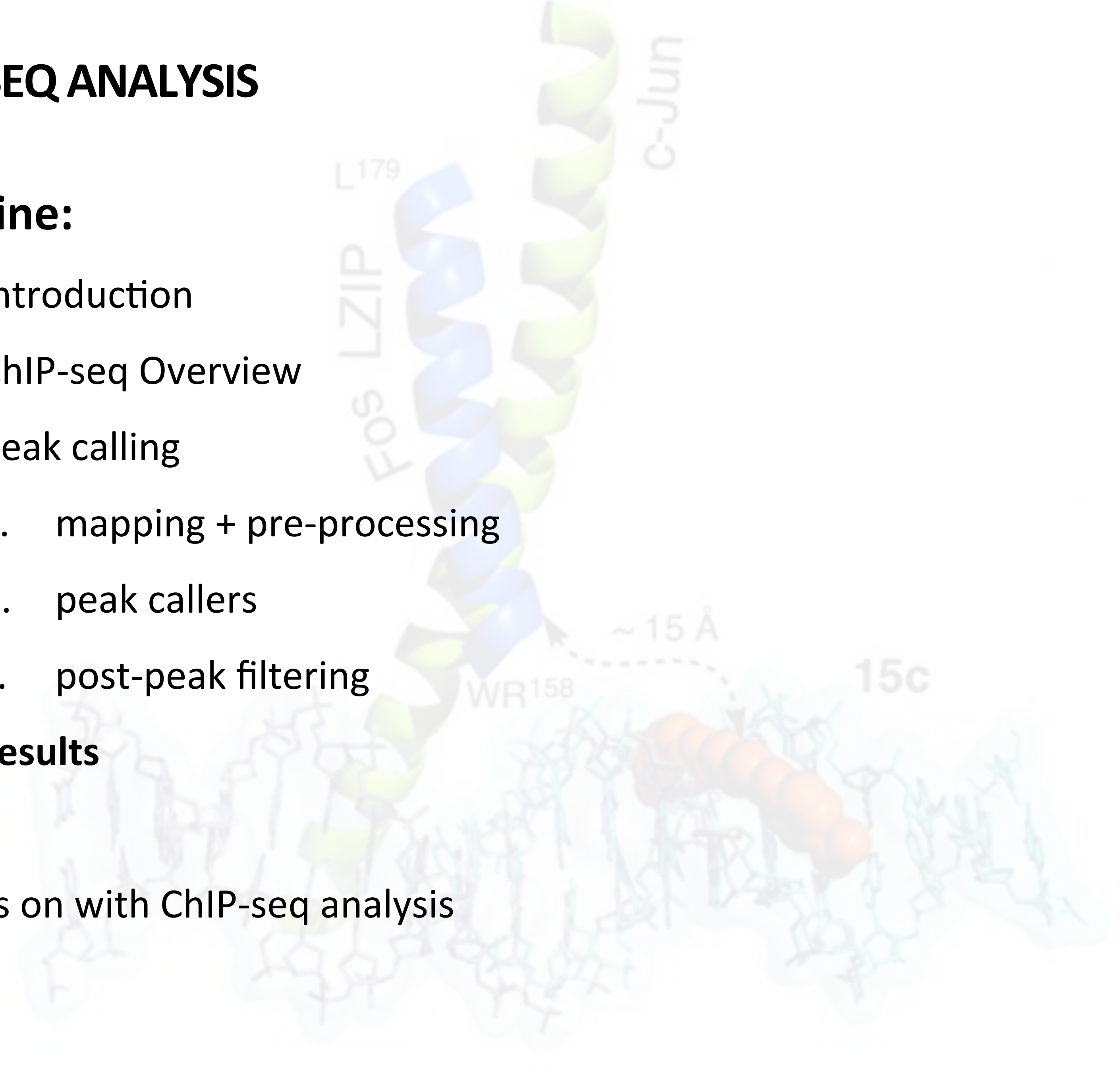
Filter peaks against the ENCODE blacklist regions

ChIP-SEQ ANALYSIS

Outline:

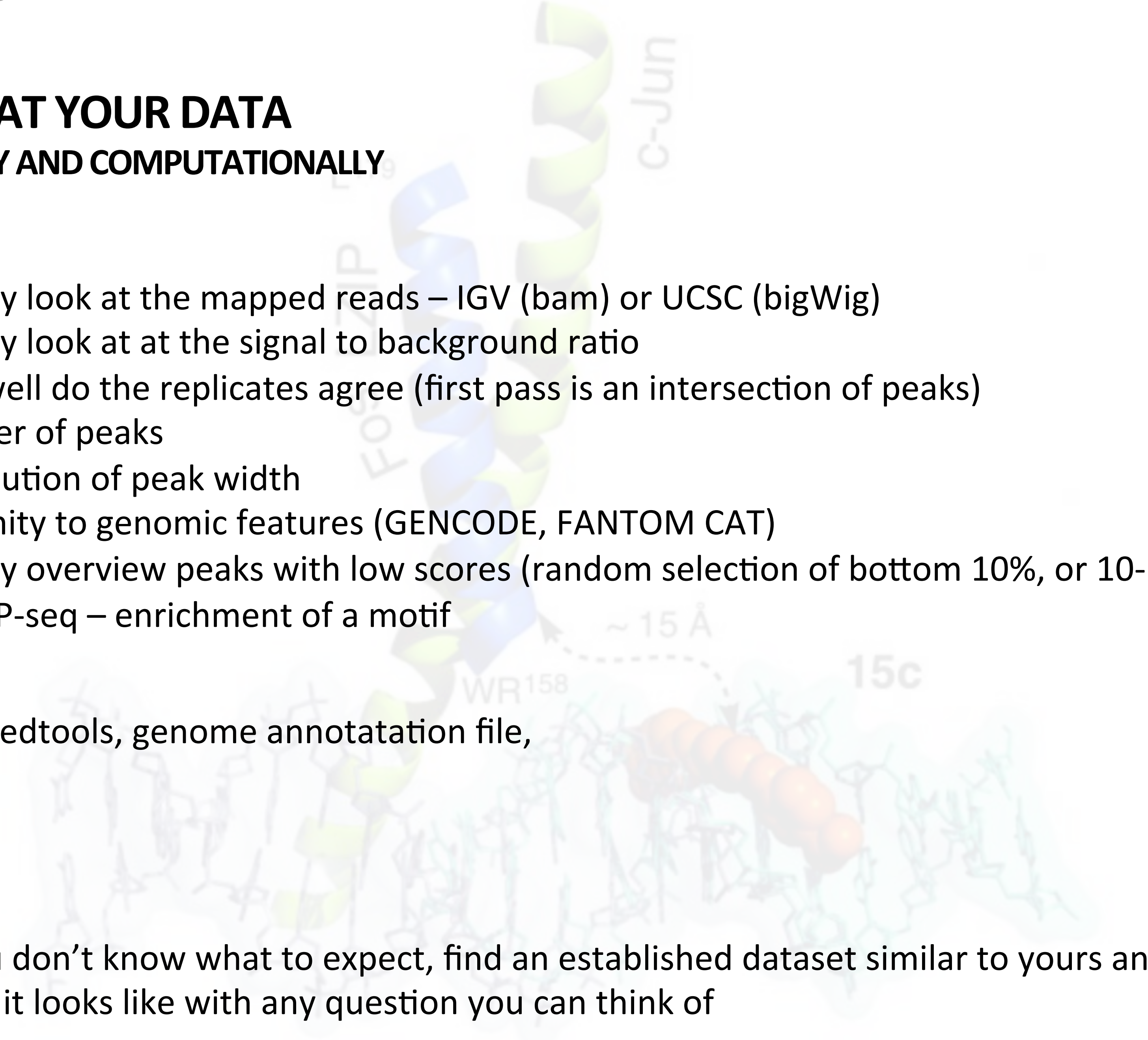
1. Introduction
2. ChIP-seq Overview
3. Peak calling
   a. mapping + pre-processing
   b. peak callers
   c. post-peak filtering
4. **Results**

Hands on with ChIP-seq analysis



LOOK AT YOUR DATA

VISUALLY AND COMPUTATIONALLY



Visually look at the mapped reads – IGV (bam) or UCSC (bigWig)
Visually look at at the signal to background ratio
How well do the replicates agree (first pass is an intersection of peaks)
Number of peaks
Distribution of peak width
Proximity to genomic features (GENCODE, FANTOM CAT)
Visually overview peaks with low scores (random selection of bottom 10%, or 10-20%)
TF ChIP-seq – enrichment of a motif

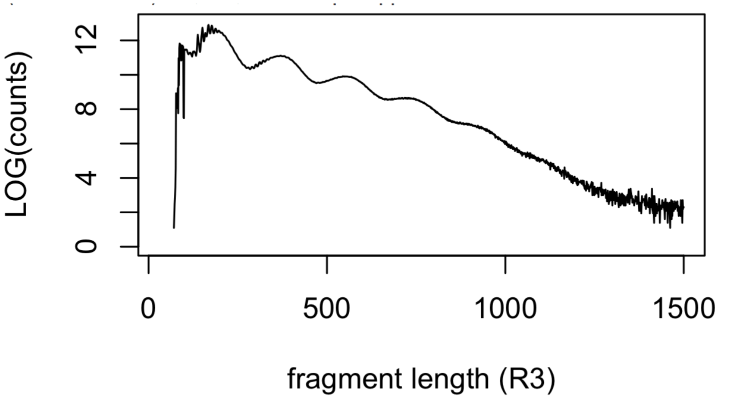
IGV, bedtools, genome annotatation file,

If you don't know what to expect, find an established dataset similar to yours and see what it looks like with any question you can think of

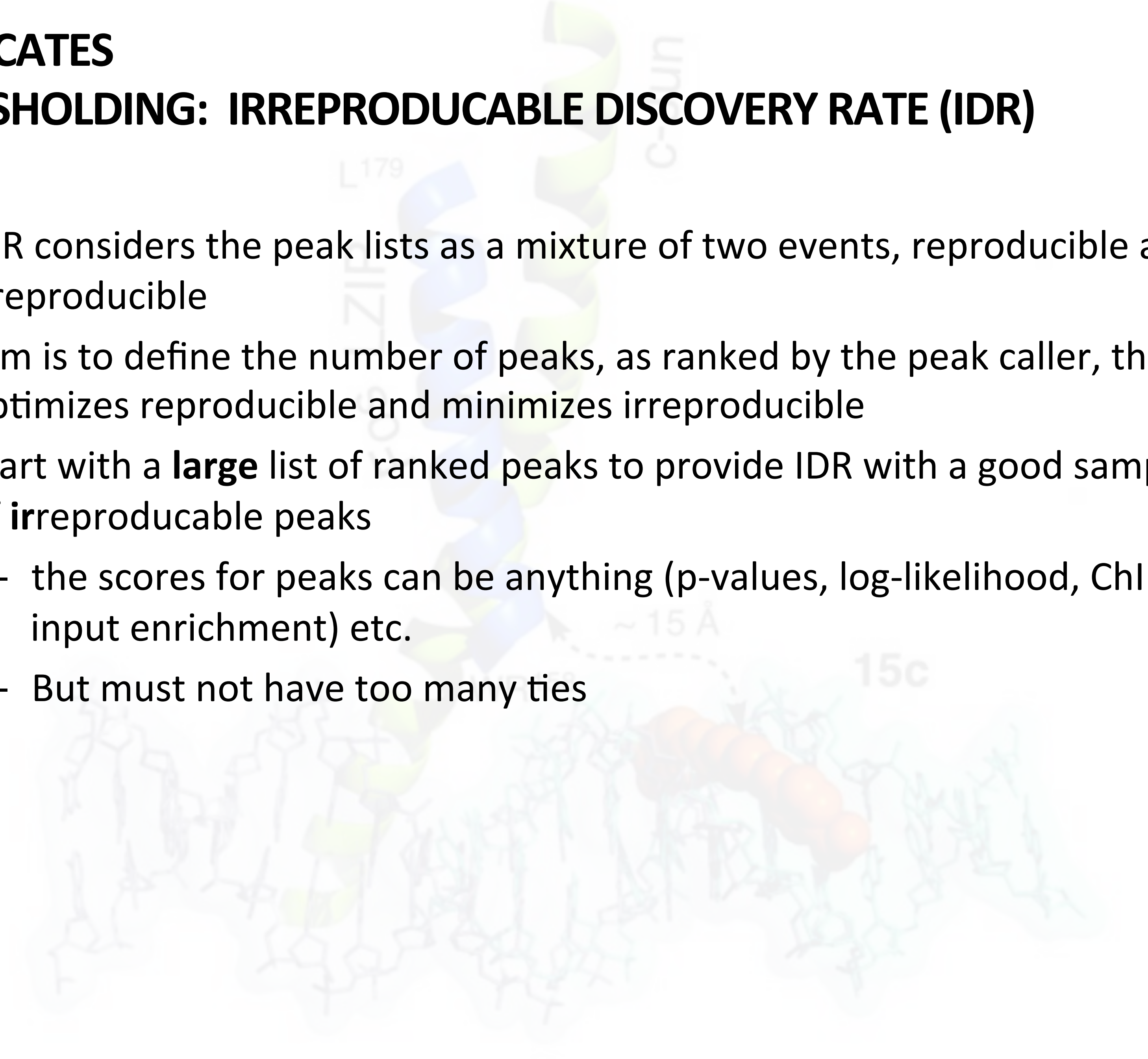
LOOK AT YOUR DATA
VISUALLY AND COMPUTATIONALLY

DNase-seq and ATAC-seq:
At least 50-60,000 peaks
~33% of peaks within 500bp of TSS (Gencode)
Peak widths: broader peaks at TSS, narrower peaks distally

Only ATAC-seq:
Look for sinusoidal pattern
~ 1:2 or 2:1 for short to long fragments



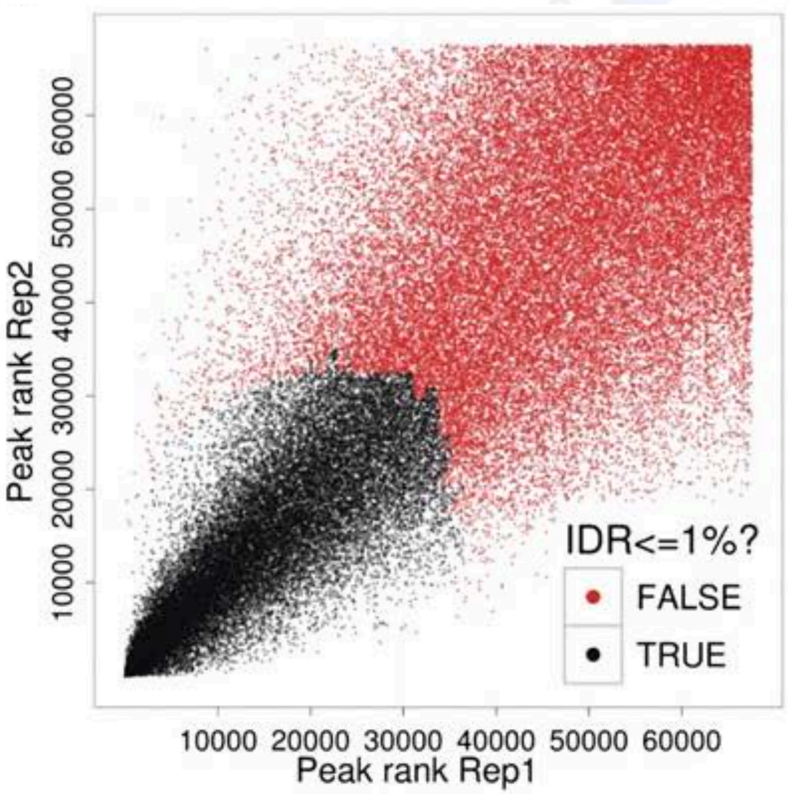
REPLICATES
THRESHOLDING: IRREPRODUCABLE DISCOVERY RATE (IDR)

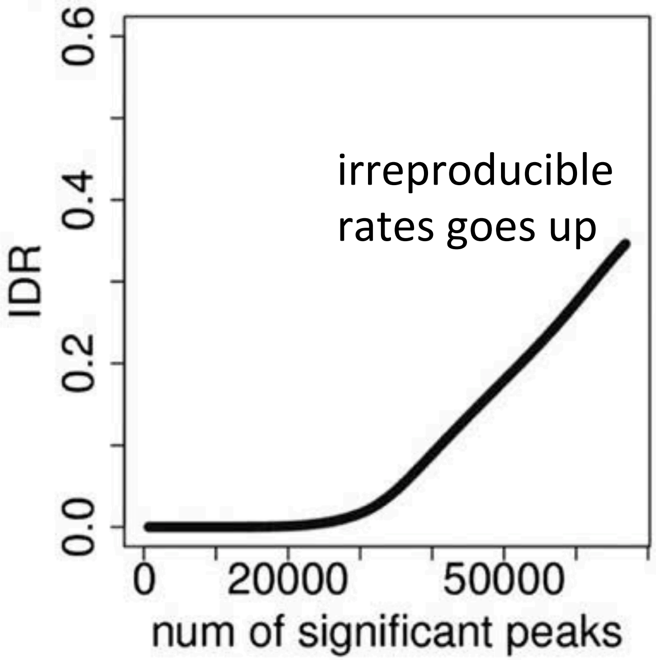


- IDR considers the peak lists as a mixture of two events, reproducible and irreproducible
- Aim is to define the number of peaks, as ranked by the peak caller, that optimizes reproducible and minimizes irreproducible
- Start with a **large** list of ranked peaks to provide IDR with a good sample of **ir**reproducable peaks
  - the scores for peaks can be anything (p-values, log-likelihood, ChIP to input enrichment) etc.
  - But must not have too many ties

IDR – UTILIZING REPLICATES

goal is to limit the expected proportion of peaks that are not reproducible across replicates

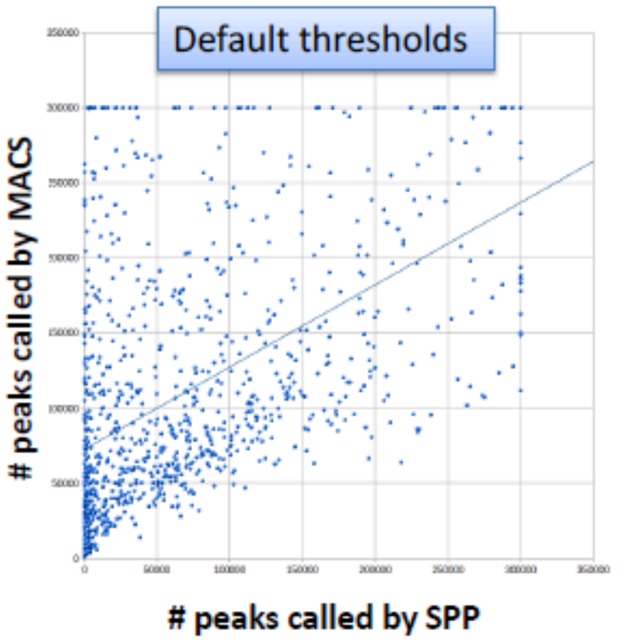


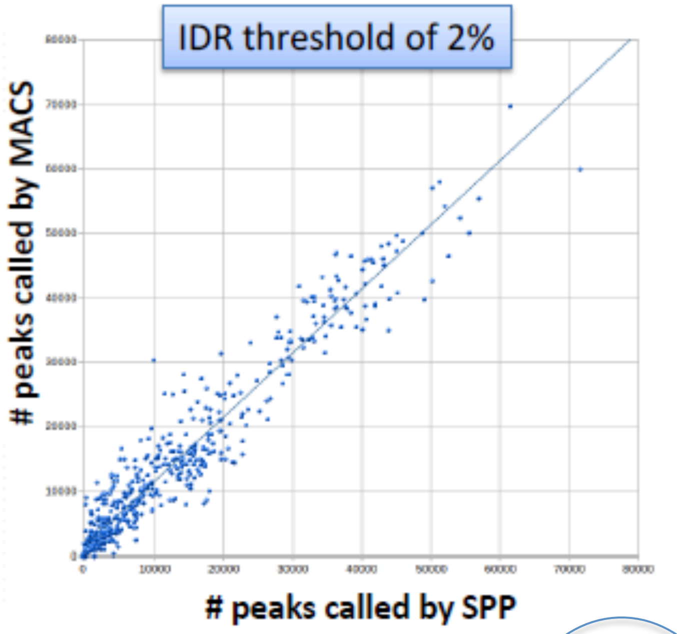


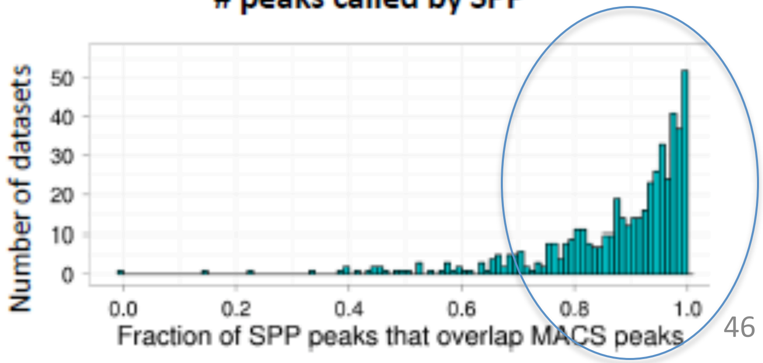
irreproducible rates goes up

IDR – STRINGENT BUT MORE CONFIDENT

Peak callers can have wildly different numbers of peak calls
After IDR different callers usually have similar numbers of peaks.



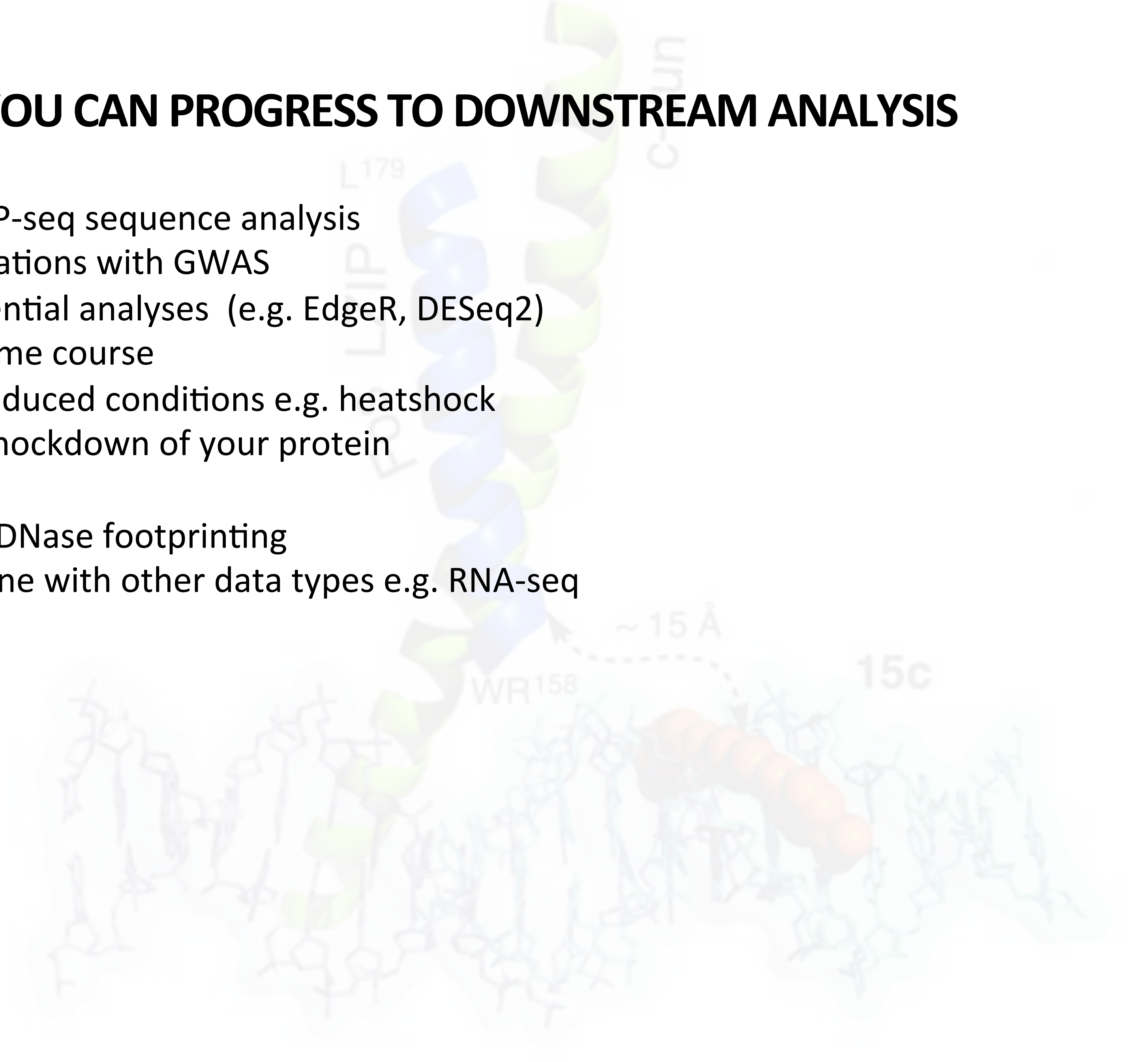




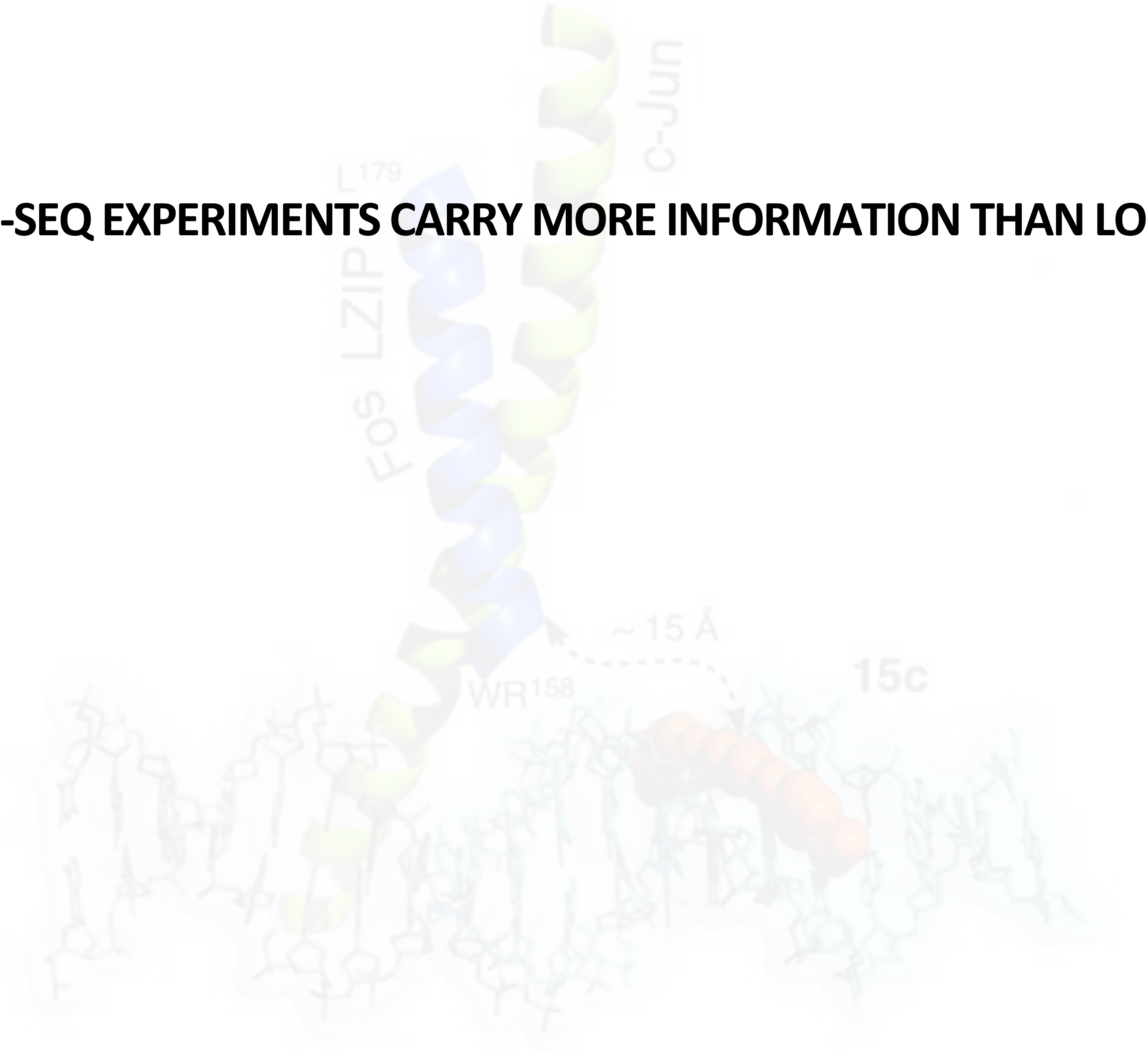
NOW YOU CAN PROGRESS TO DOWNSTREAM ANALYSIS

TF ChIP-seq sequence analysis
Associations with GWAS
Differential analyses (e.g. EdgeR, DESeq2)
    time course
    induced conditions e.g. heatshock
    knockdown of your protein

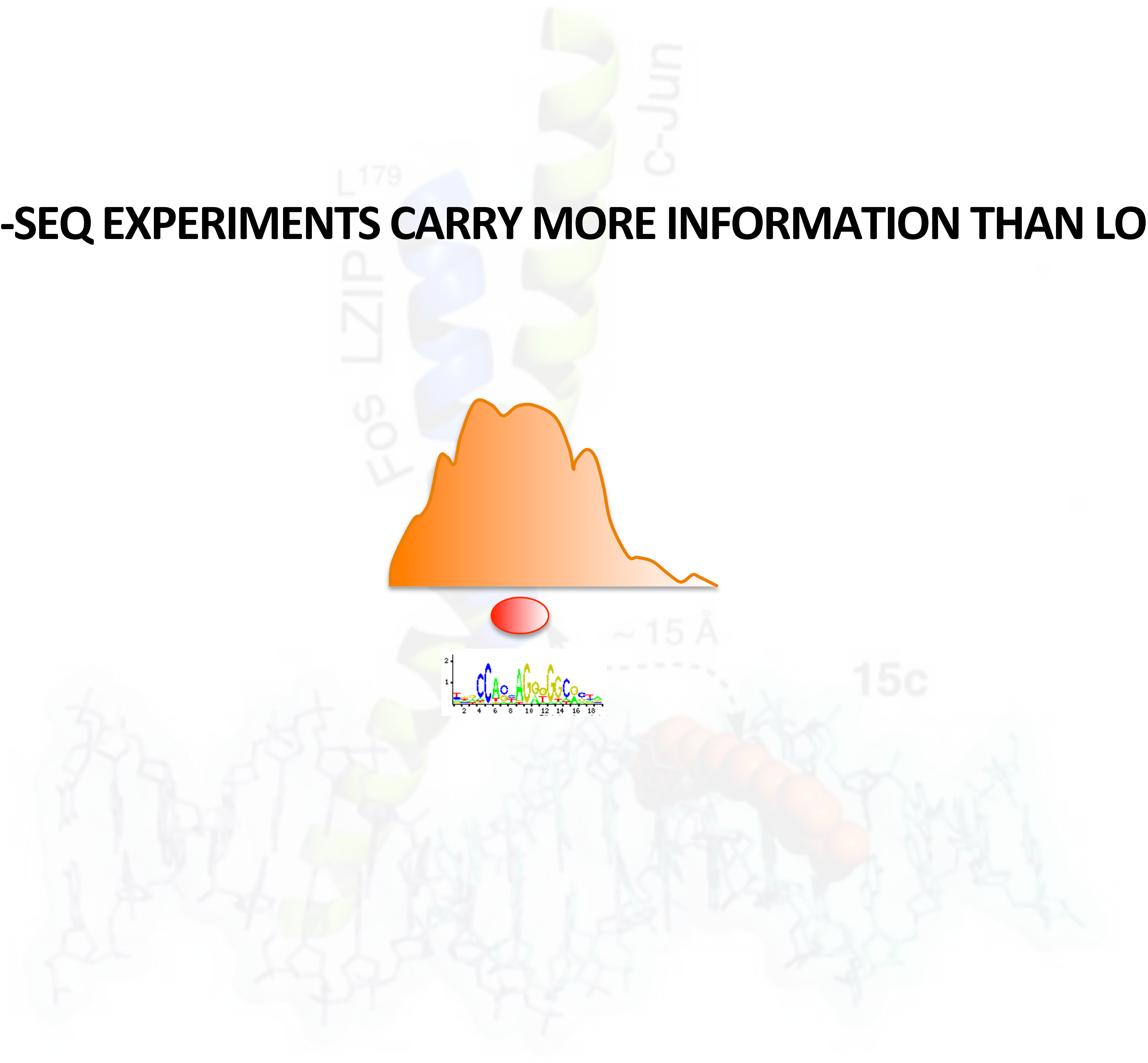
ATAC/DNase footprinting
Combine with other data types e.g. RNA-seq



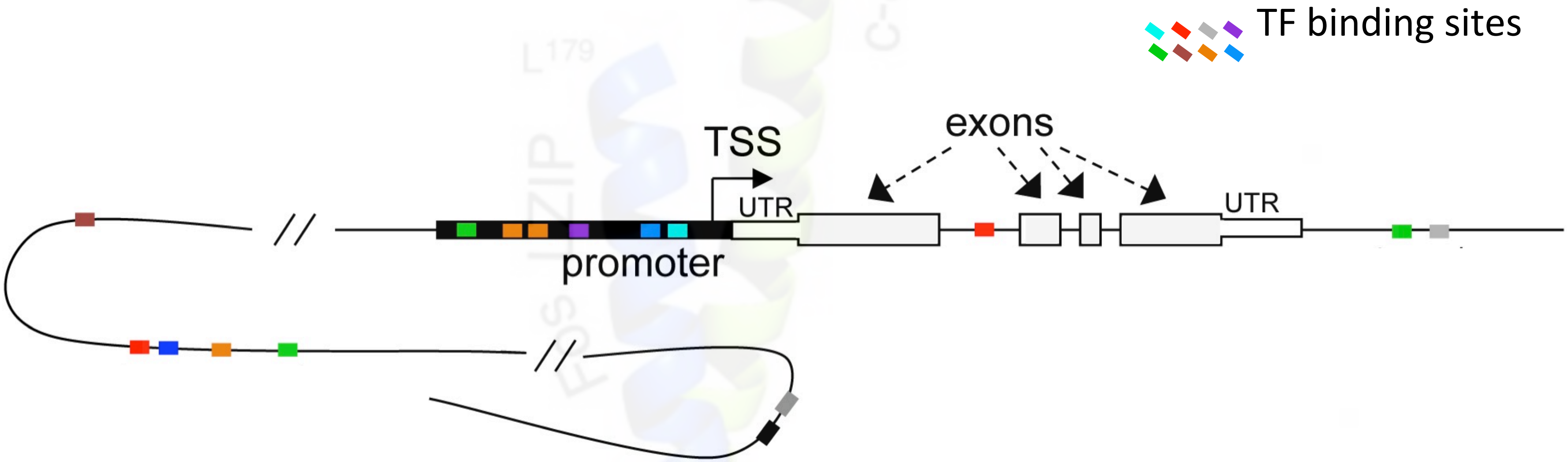
TF ChIP-SEQ EXPERIMENTS CARRY MORE INFORMATION THAN LOCATION

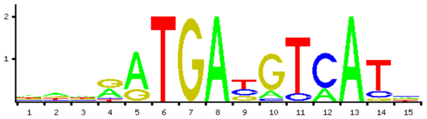


TF ChIP-SEQ EXPERIMENTS CARRY MORE INFORMATION THAN LOCATION



TFs RECOGNIZE SEQUENCE PATTERNS





....gggGA**TGA**CG**TCA**Tc....
....gaaAA**TGA**TG**CAA**Ca....
....tggAG**TGA**TG**CAA**Ta....
....ttcAG**TGA**GA**TCA**Cc....
....ggtGA**TGA**TG**TCA**Tc....
....aatAG**TGA**TG**CAA**Ta....
....agcGA**TGA**CG**TCA**Tc....
....gaaGA**TGA**TG**TCA**Tc....

...

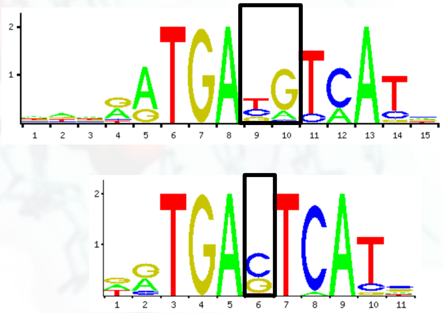
...

...

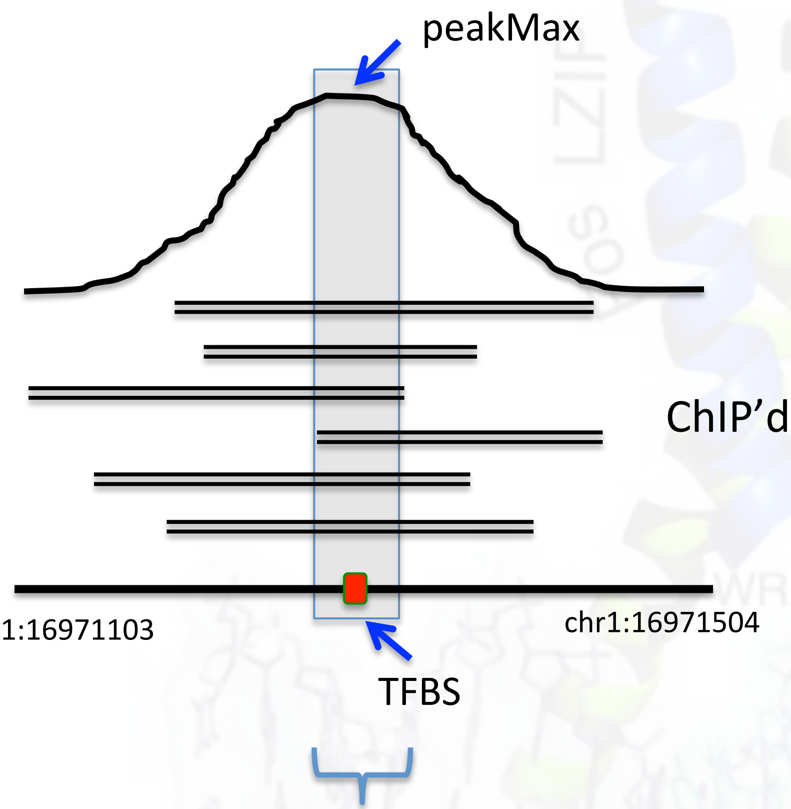
...

....ggtGA**TGA**AG**TCA**Cc....

JunD alternative spacing



ENRICHMENT OF KNOWN MOTIFS WITH RESPECT TO THE PEAK MAX/ SUMMIT



peakMax

ChIP’d sequence fragments

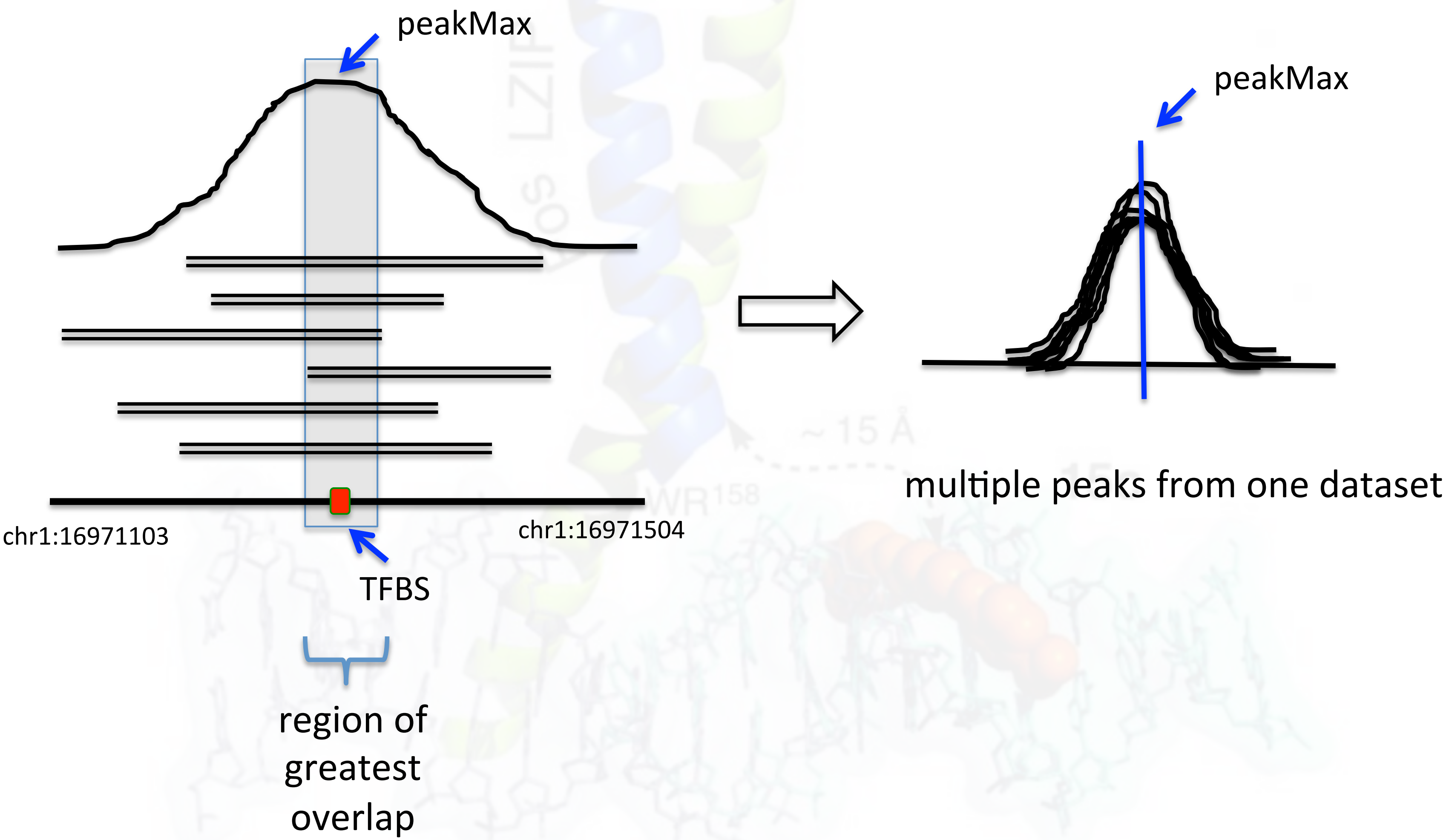
chr1:16971103

chr1:16971504

TFBS

region of greatest overlap

ENRICHMENT OF KNOWN MOTIFS WITH RESPECT TO THE PEAK MAX/ SUMMIT



peakMax

chr1:16971103

chr1:16971504

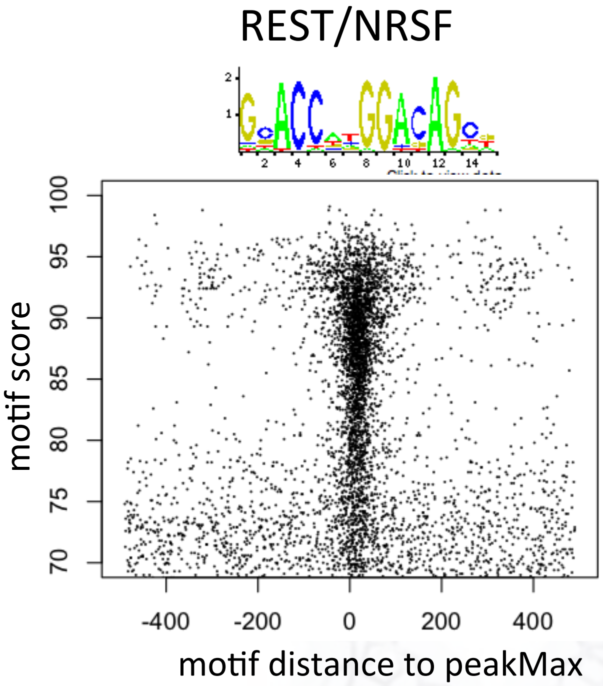
TFBS

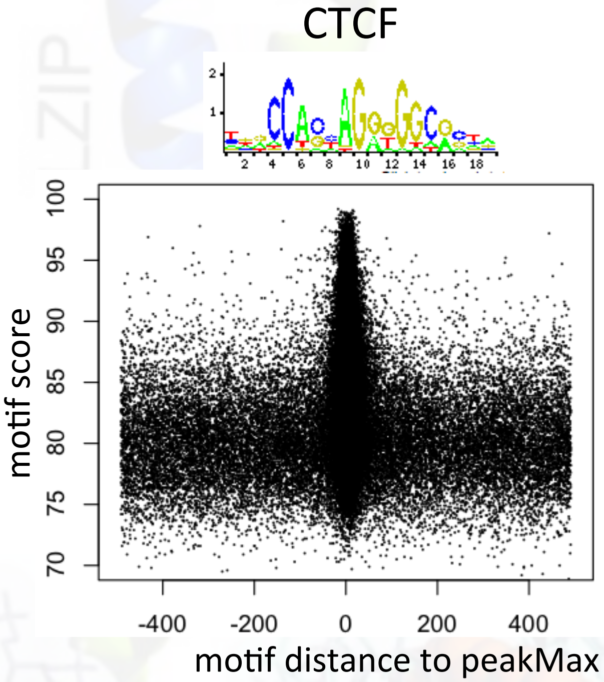
region of greatest overlap

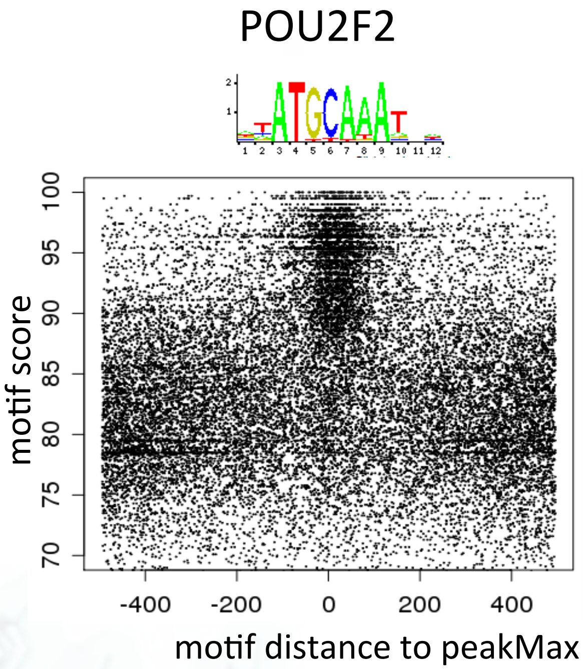
peakMax

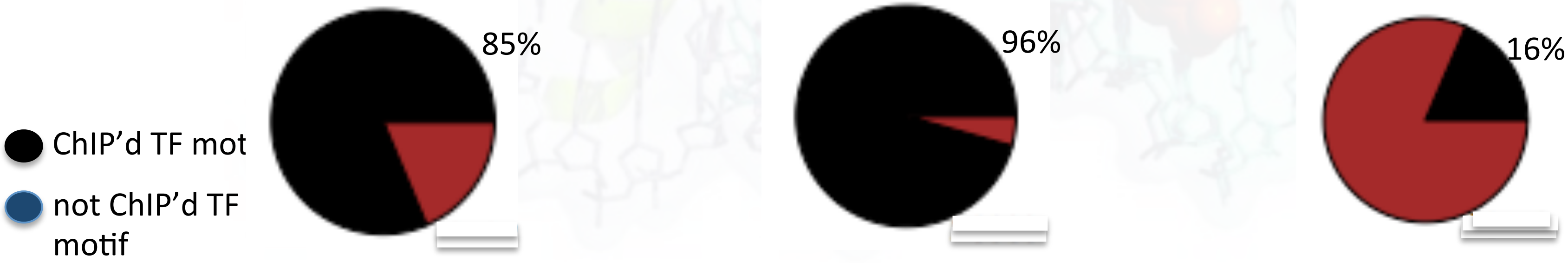
multiple peaks from one dataset

ENRICHMENT OF MOTIFS WITH RESPECT TO THE PEAKMAX IS SUPPORTING EVIDENCE OF DIRECT BINDING





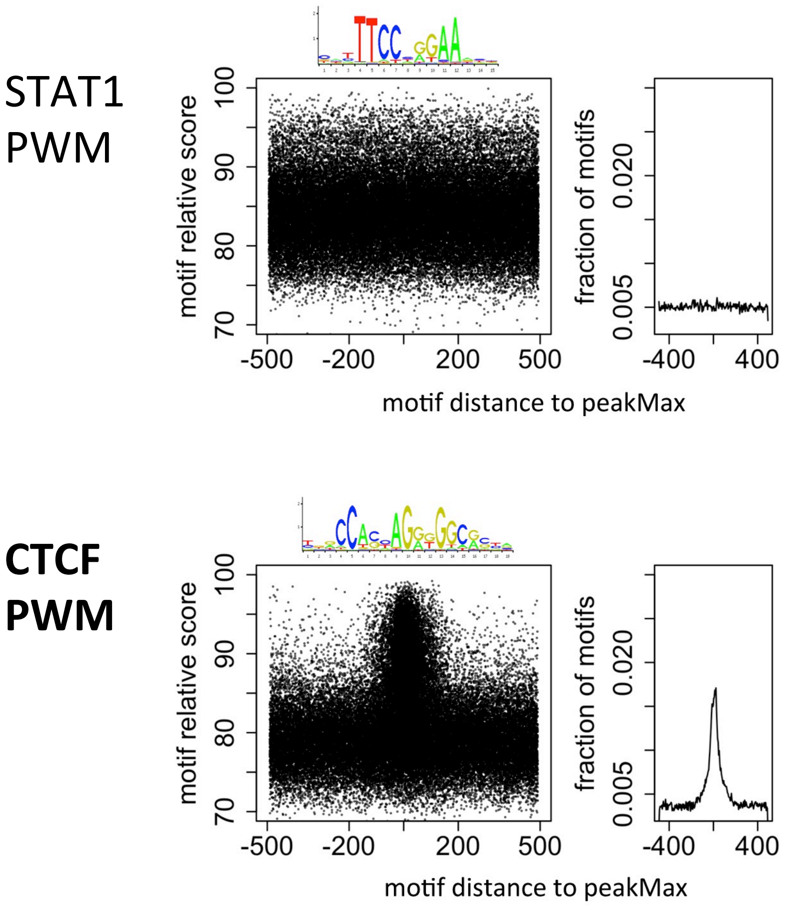




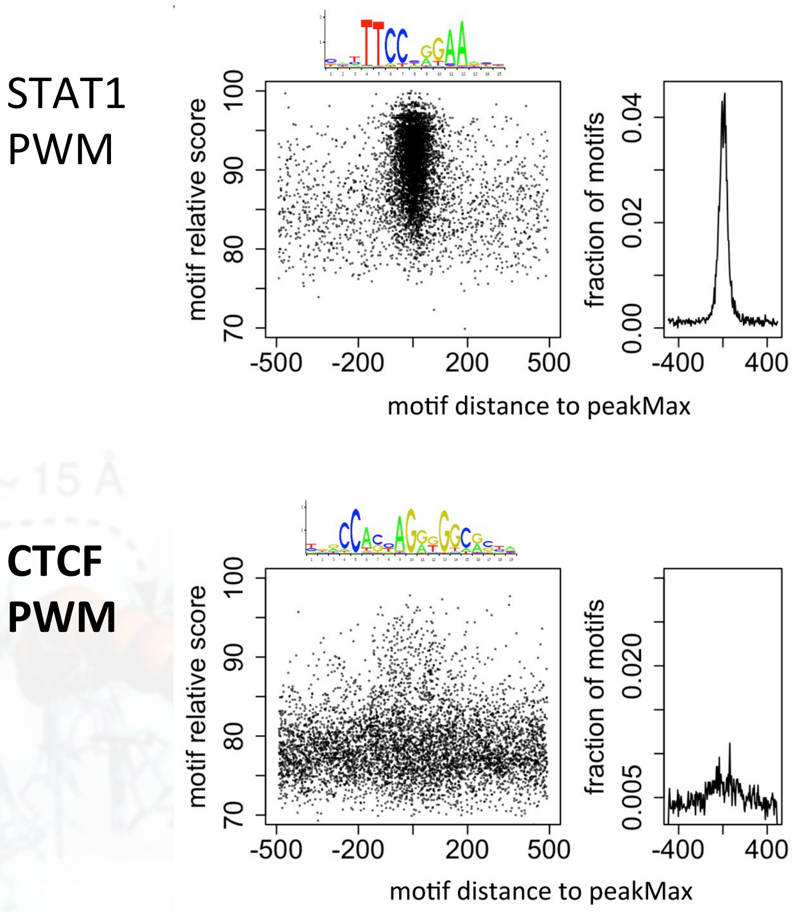
Worsley Hunt and Wasserman. Genome Biol. 2014

STAT1 - AN EXTREME EXAMPLE OF NON-TARGETED TF MOTIF ENRICHMENT

Stat1 GM12878 ChIP-seq



Stat1 HeLa IFN-γ ChIP-seq



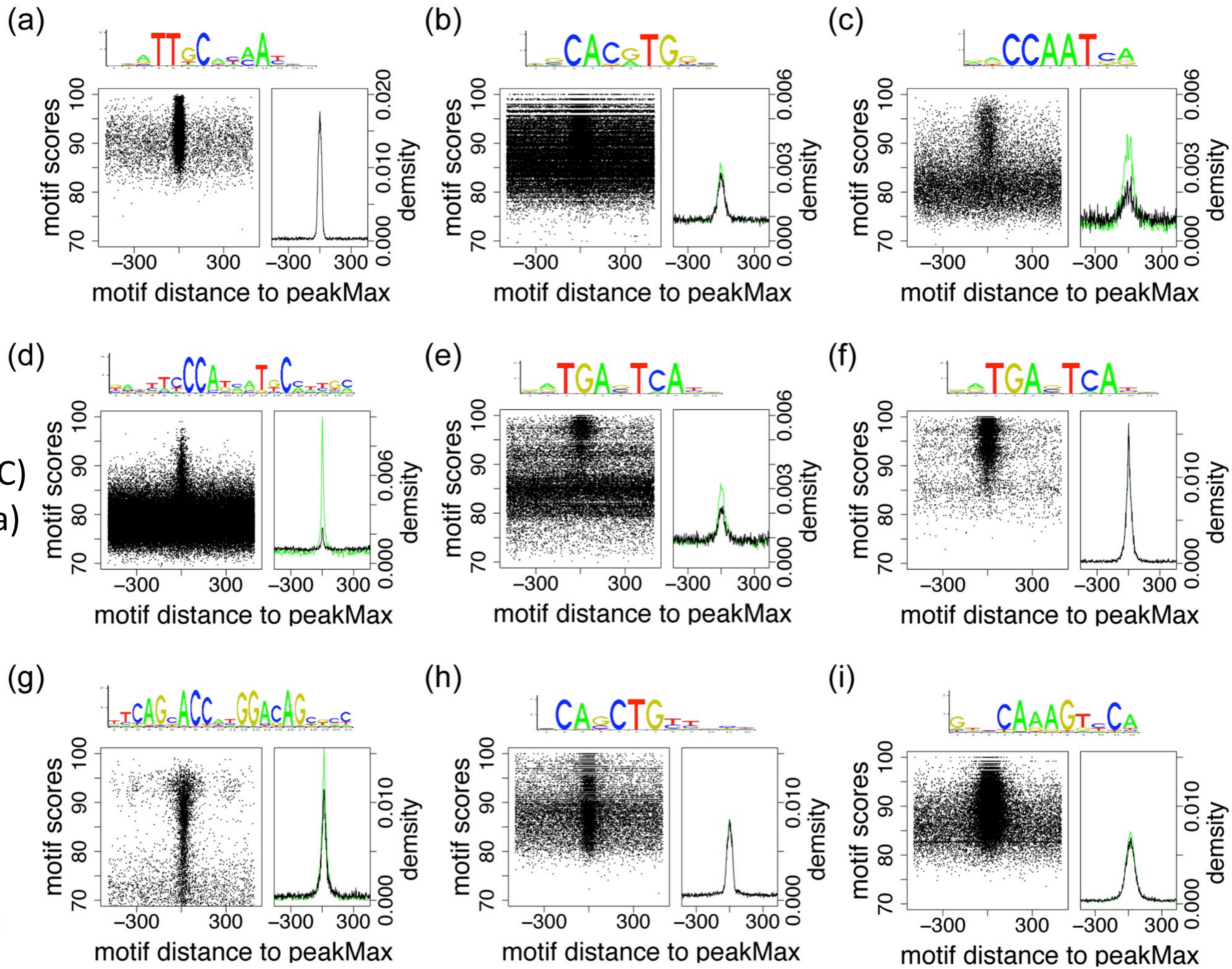
Worsley Hunt and Wasserman. Genome Biol. 2014

4/4 Results

a) C/Ebp-B
b) cMYC
c) Nf-yA

d) ZNF143
e) JUN (hESC)
f) JUN (HeLa)

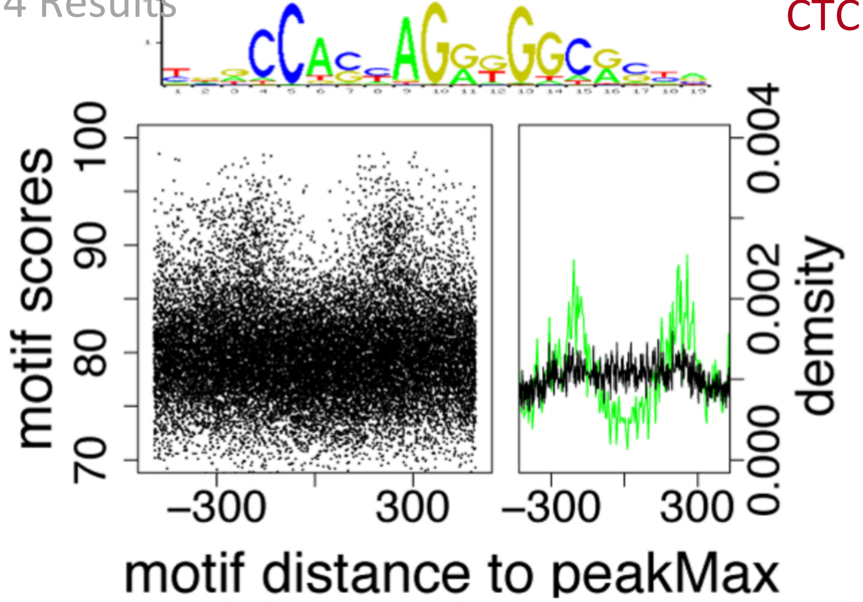
g) REST
h) MyoD
i) HNF4A



Worsley Hunt and Wasserman. Genome Biol. 2014

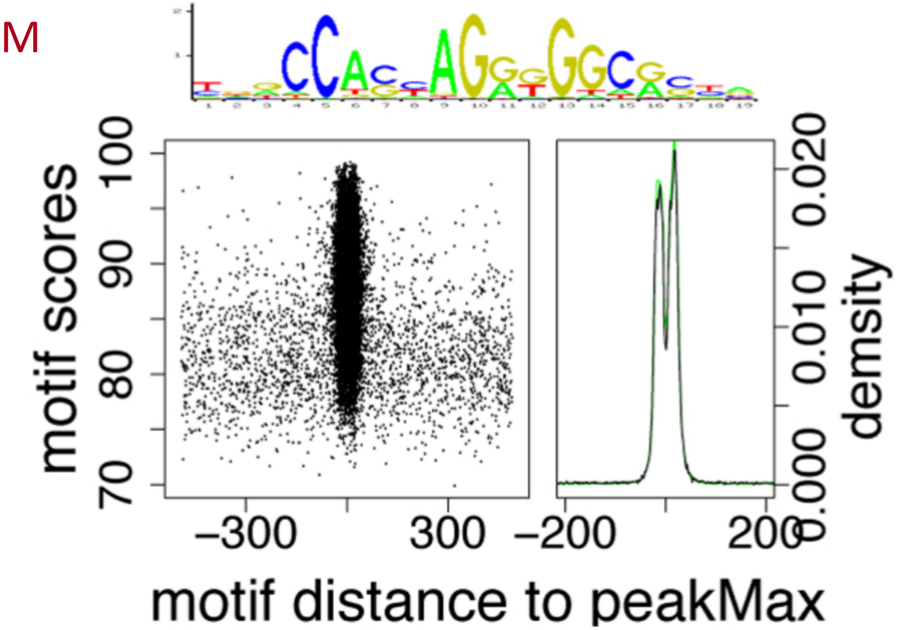
4/4 Results

CTCF PWM



0 =

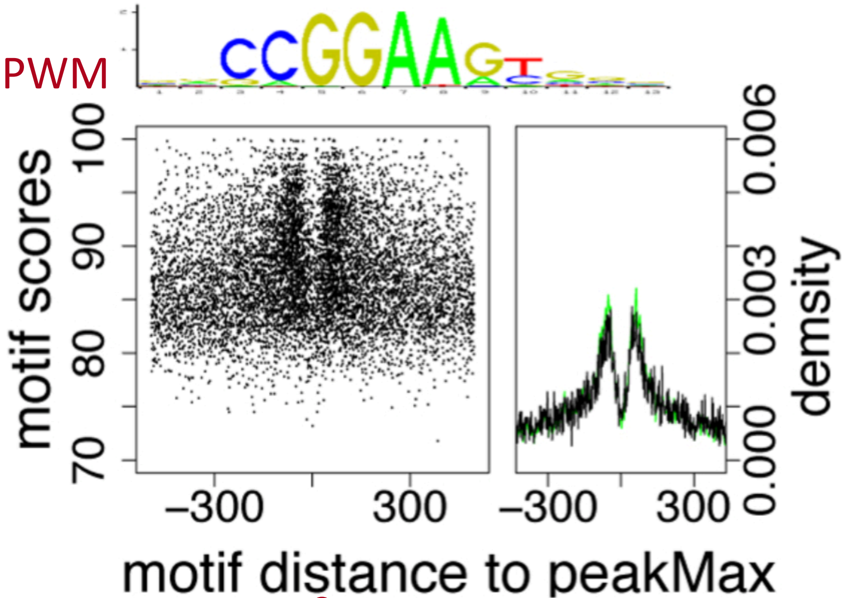
peakMax H3K4me3 ChIP-seq



0 =

peakMax RAD21::cohesin ChIP-seq

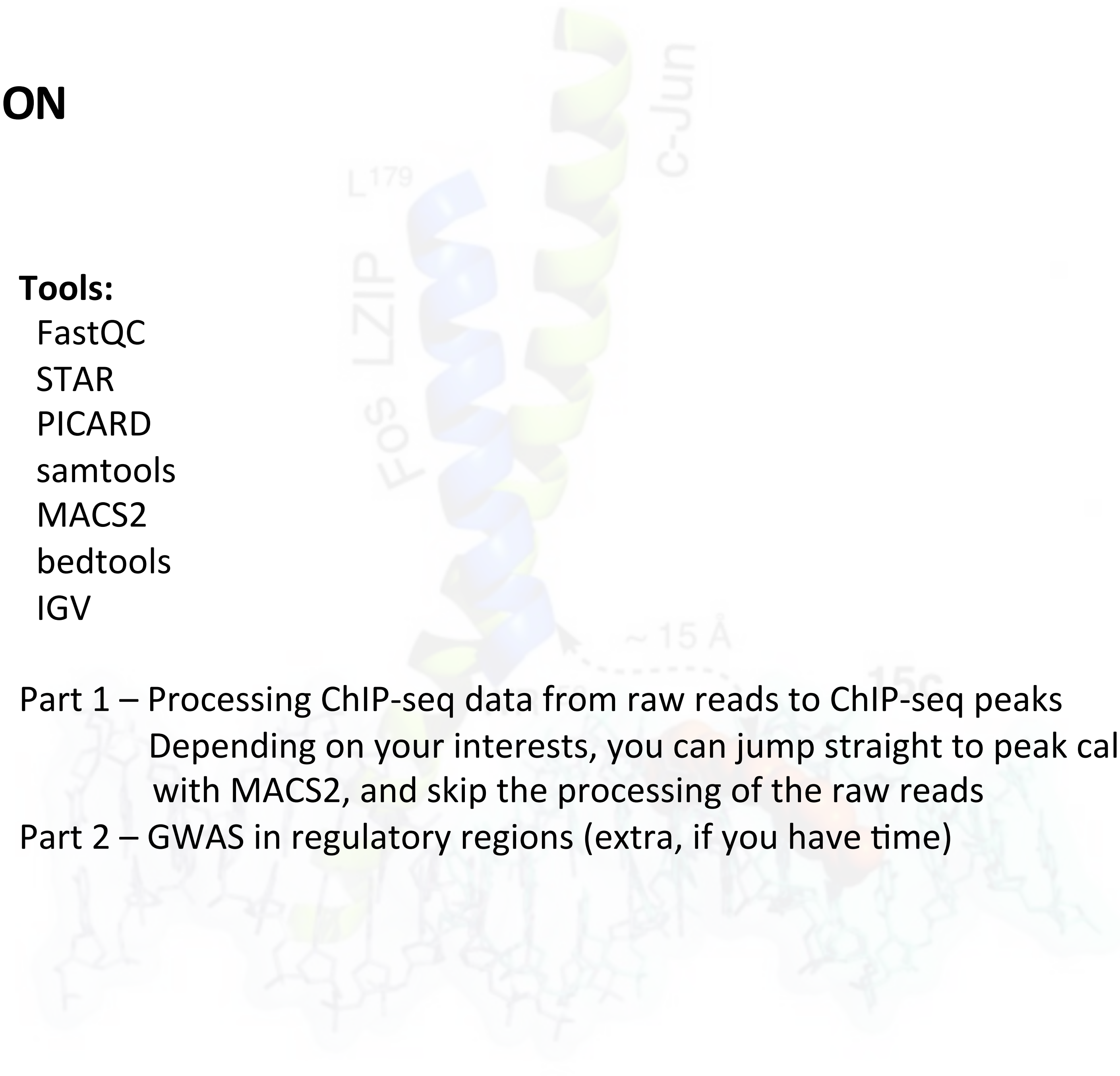
ELK4 PWM



0 =

peakMax NELFE ChIP-Seq

HANDS ON



Tools:
FastQC
STAR
PICARD
samtools
MACS2
bedtools
IGV

Part 1 – Processing ChIP-seq data from raw reads to ChIP-seq peaks
Depending on your interests, you can jump straight to peak calling with MACS2, and skip the processing of the raw reads

Part 2 – GWAS in regulatory regions (extra, if you have time)