Analysis of ChIP-seq data
BIOC8145

Chongzhi Zang

zang@virginia.edu
zanglab.org

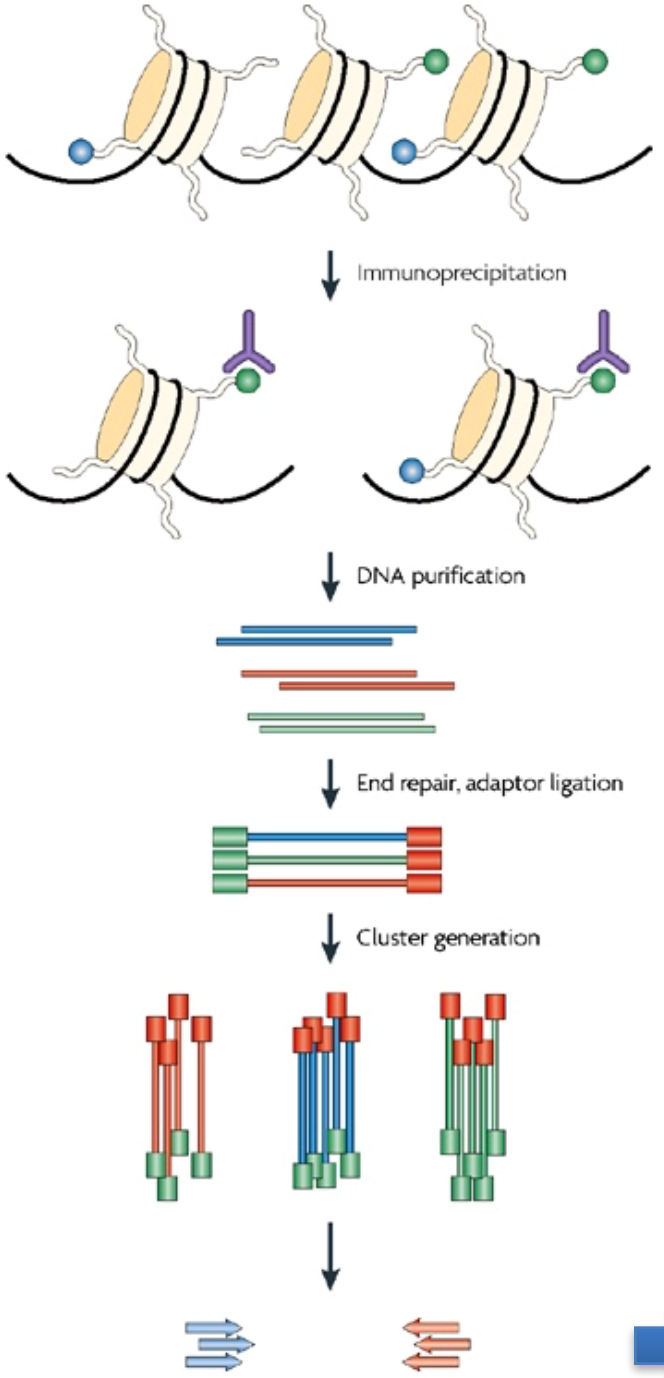
BIOC8145 – Spring 2020
April 6, 2020

Outline

- Lecture 1
  - ChIP-seq technique introduction
  - ChIP-seq data analysis strategy
  - Read mapping (bowtie2)
  - Data formats
- Lecture 2
  - Peak calling (macs2)
  - Data visualization (IGV)
  - Quality control
- Lecture 3
  - Downstream analysis and integration
  - Online resources

Lecture 2: ChIP-seq Analysis

- Data processing (continued)
- Peak calling using **macs2**
- Quality control
- Data visualization



Experimental procedure

Biology

downstream analysis/integration

Peaks (bed)

peak calling

Pile-up for visualization (bedGraph, wig, bigwig)

macs2

Non-redundant reads (sam/bam/bed)

redundancy assessment

Mapped reads (sam/bam/bed)

alignment (**bowtie2**)

Raw sequence reads (fastq)

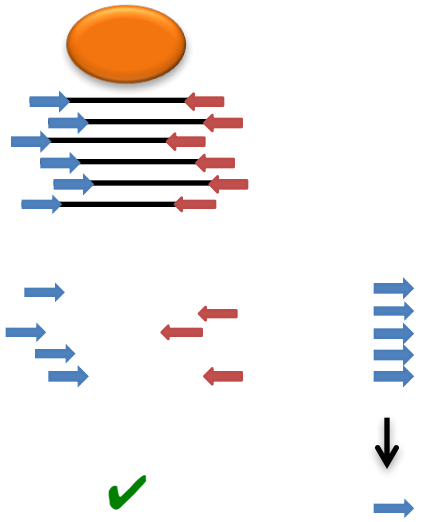
Computational analysis

ChIP-seq: Data processing

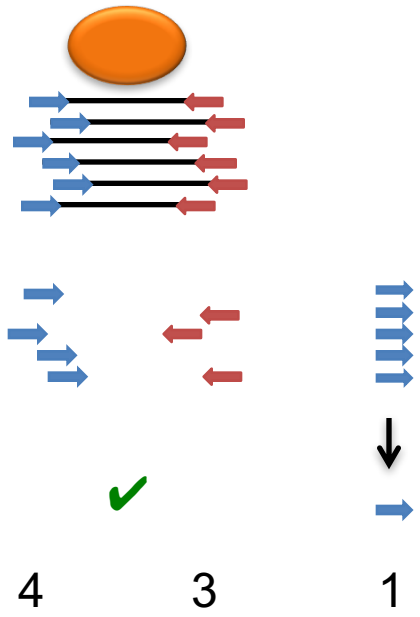
- alignment of each sequence read: **bowtie2** or **BWA**

cannot map to the reference genome ✗
can map to multiple loci in the genome ✗
can map to a unique location in the genome ✔

- redundancy control:



Redundancy Control



# mapped reads:	12
# non-redundant reads:	8
# locations w/ reads:	8
# locations w/ 1 read:	7

- Non-redundant rate:

$$\frac{\text{\# non-redundant reads}}{\text{\# mapped reads}}$$

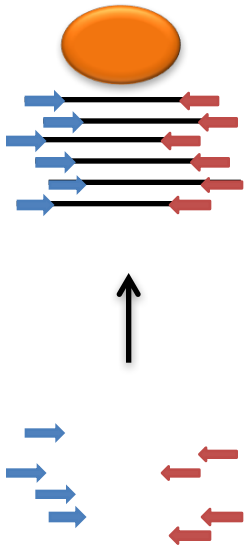
8/12 = 66.7%

- PBC (PCR Bottleneck Coefficient):

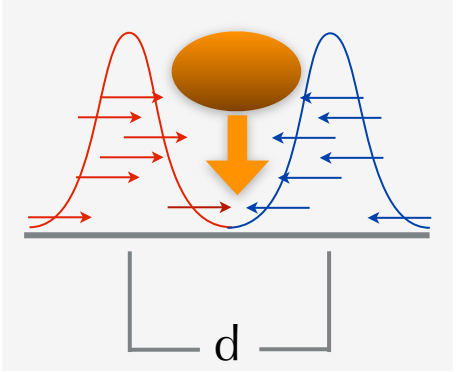
$$\frac{\text{\# locations w/ 1 read}}{\text{\# locations w/ reads}}$$

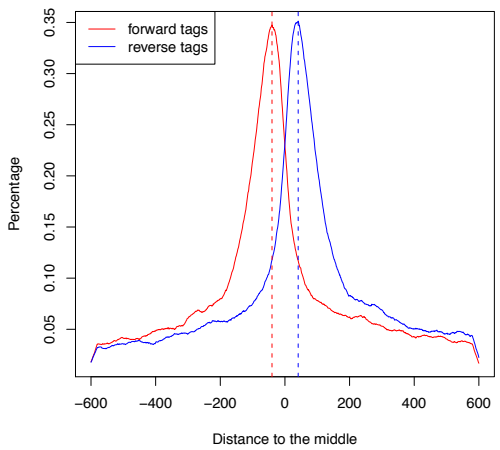
7/8 = 87.5%

DNA fragment size estimation

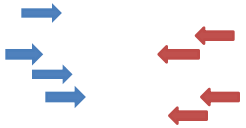


peak model

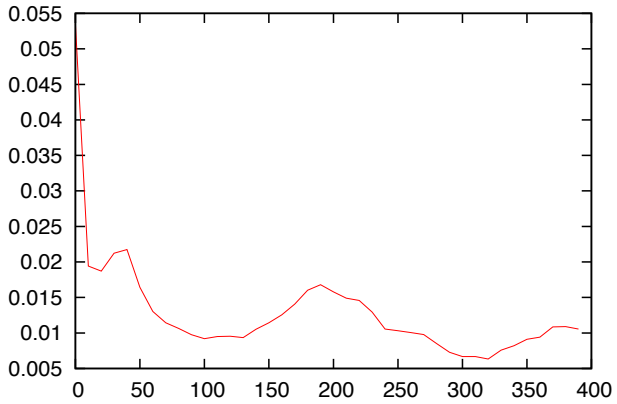




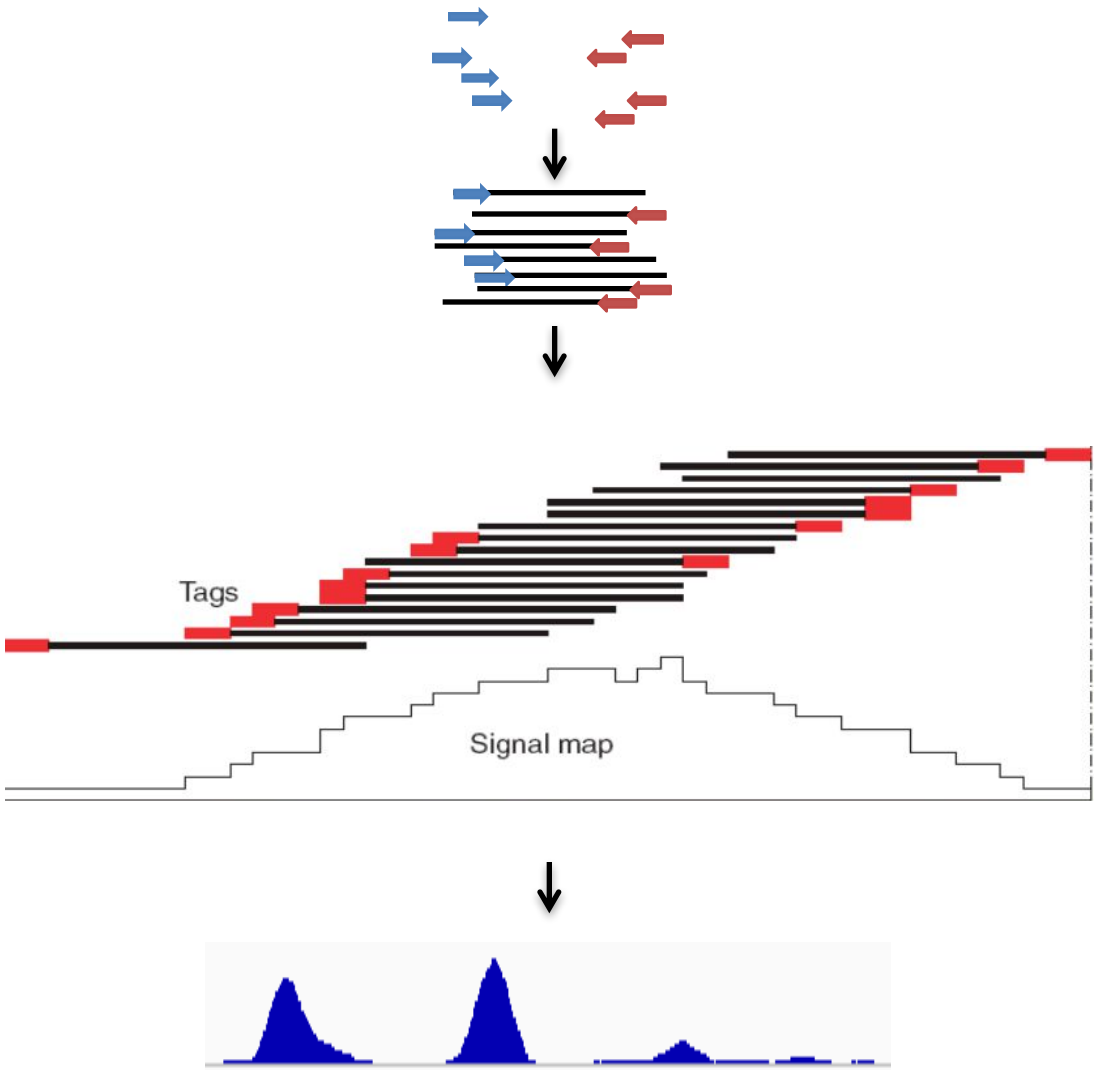
cross-correlation



$$C(r) = \frac{1}{X} \int_x \left(T_+(x) - \overline{T_+}\right) \left(T_-(x+r) - \overline{T_-}\right)$$



Pile up: visualization



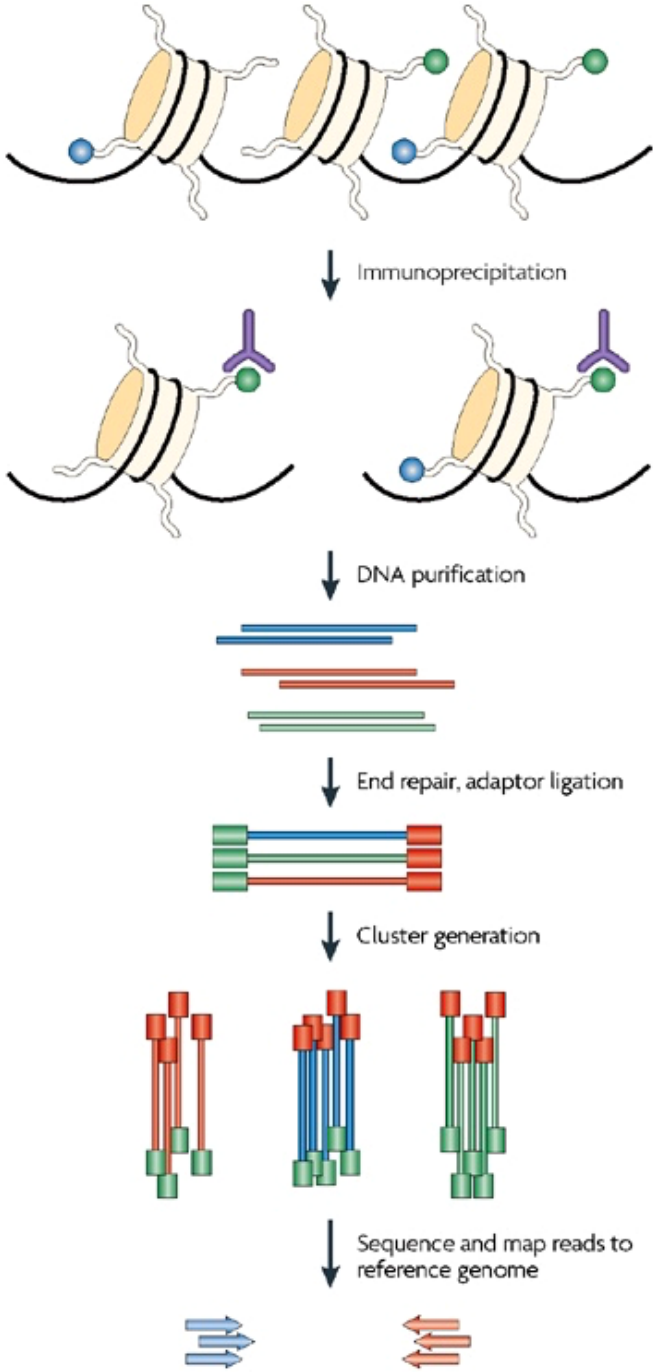
- bedGraph:

```
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



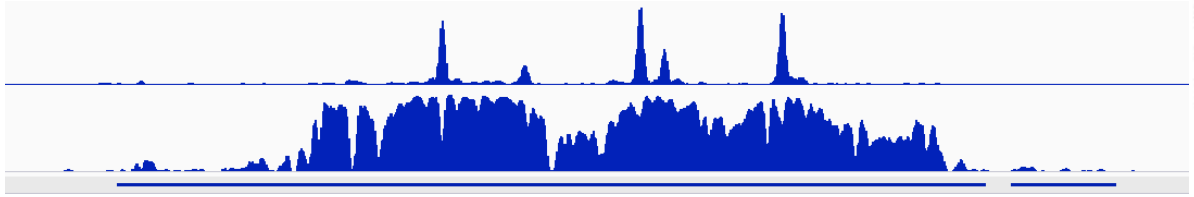
ChIP-seq: Study design

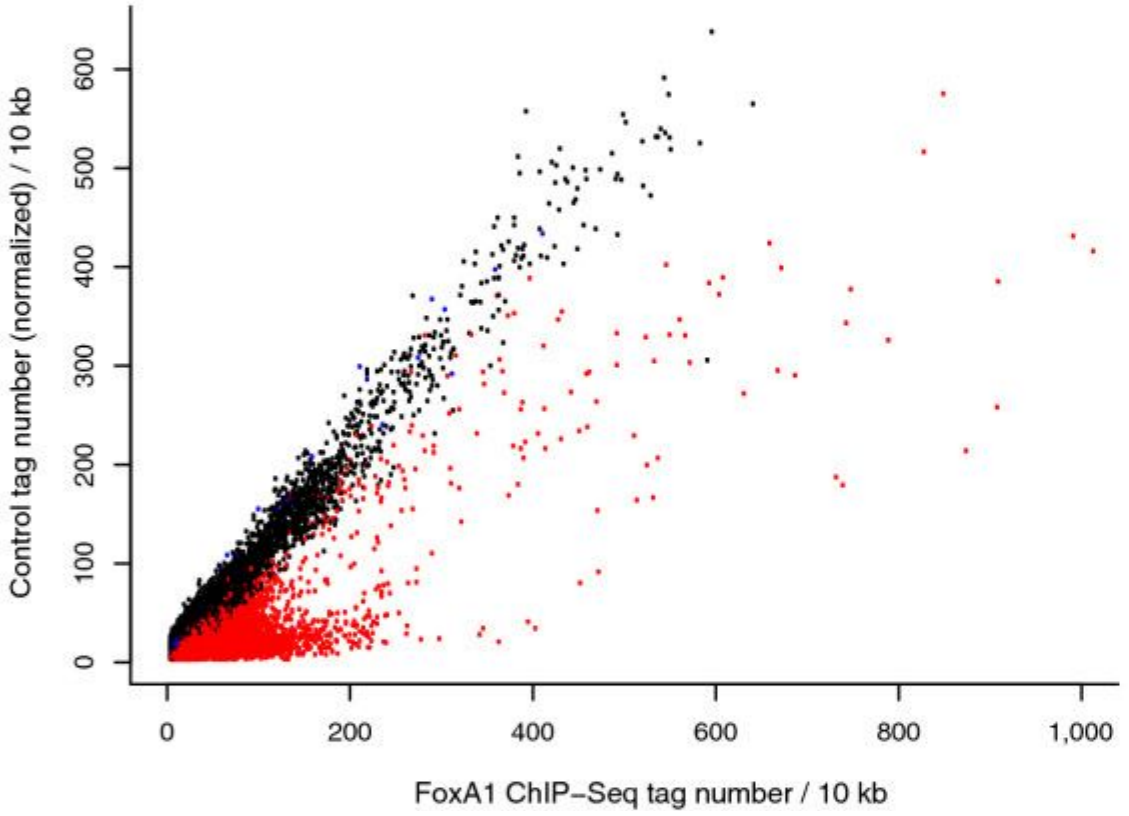
- Background Control: **Input** or **IgG**
  - Input chromatin: sonicated/digested chromatin without immunoprecipitation
  - IgG: “unspecific” immunoprecipitation

- Study Control:
  - Control exp sample: ChIP + input
  - Treated exp sample: ChIP + input

ChIP-seq: Peak calling

- Goal: Identify regions in the genome enriched for sequence reads:
  - Compared to genomic background
  - Compared to input control



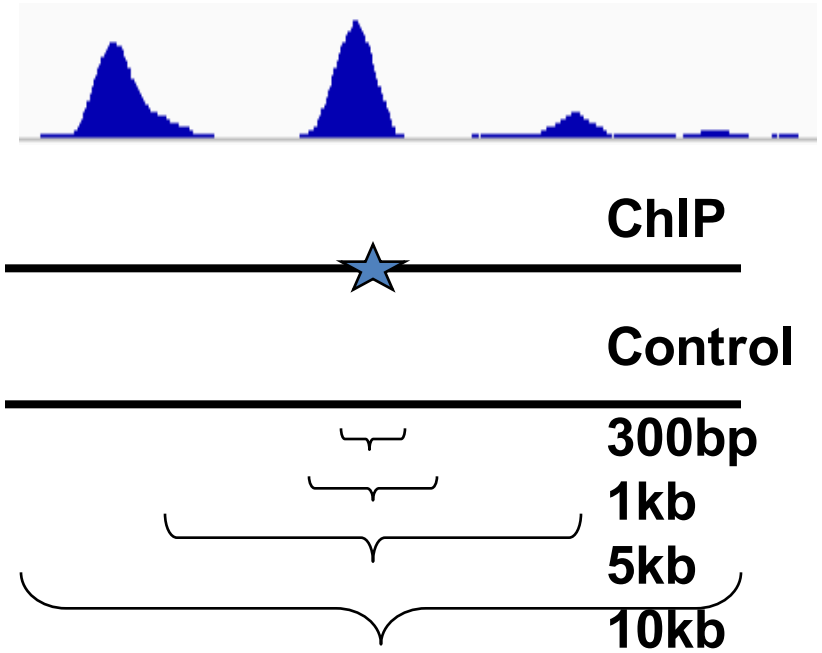


MACS: model

- **M**odel-based **A**nalysis for **C**hIP-**S**eq
- Read distribution along the genome ~ Poisson distribution
  (λ_{BG} = total tag / genome size)
- ChIP-seq show local biases in the genome
  - Chromatin and sequencing bias
  - 200-300bp control windows have to few tags
  - But can look further

Dynamic λ_{local} =
max(λ_{BG}, [λ_{ctrl}, λ_{1k},] λ_{5k}, λ_{10k})

- B-H adjustment to correct for FDR
  - p-value → q-value



Zhang et al, *Genome Bio*, 2008

MACS: Critical input parameters

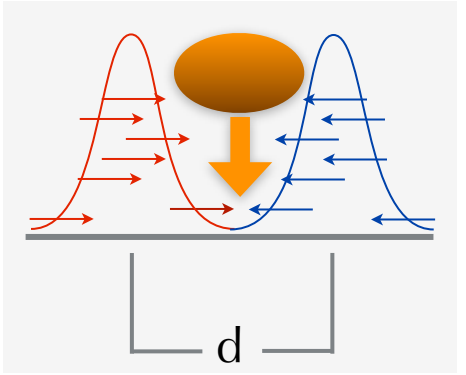
```
macs2 callpeak [-h] -t TFILE [TFILE ...] [-c [CFILE]] [-g GSIZE] [-q QVALUE | -p PVALUE] [--
outdir OUTDIR] [-n NAME] [-B]

-g GSIZE              Effective genome size. It can be 1.0e+9 or 1000000000,
                      or shortcuts:'hs' for human (2.7e9), 'mm' for mouse
                      (1.87e9), 'ce' for C. elegans (9e7) and 'dm' for
                      fruitfly (1.2e8), Default:hs
-q QVALUE
                      Minimum FDR (q-value) cutoff for peak detection.
                      DEFAULT: 0.05. -q, and -p are mutually exclusive.

--outdir OUTDIR       If specified all output files will be written to that
                      directory. Default: the current working directory

-n NAME               Experiment name, which will be used to generate output
                      file names. DEFAULT: "NA"

-B, --bdg             Whether or not to save extended fragment pileup, and
                      local lambda tracks (two files) at every bp into a
                      bedGraph file. DEFAULT: False
```

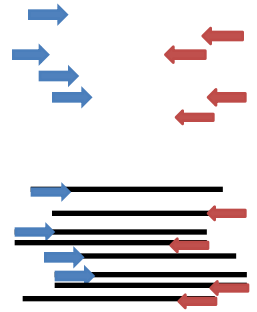
MACS: Output interpretation

```
# This file is generated by MACS version 2.1.2
# Command line: callpeak -t ../bowtie2/AR.sam -g hs -n AR --bdg
# ARGUMENTS LIST:
# name = AR
# format = AUTO
# ChIP-seq file = ['../bowtie2/AR.sam']
# control file = None
# effective genome size = 2.70e+09
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-02
# The maximum gap between significant sites is assigned as the read length/tag size.
# The minimum length of peaks is assigned as the predicted fragment length "d".
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 10000 bps
# Broad region calling is off
# Paired-End mode is off
```

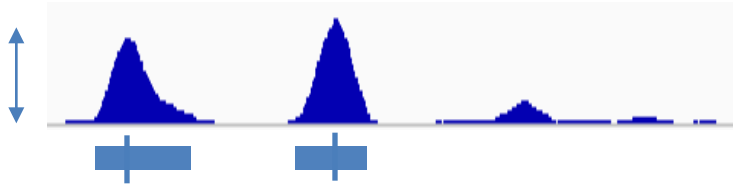
MACS: Output interpretation

```
# tag size is determined as 51 bps
# total tags in treatment: 19442622
# tags after filtering in treatment: 17218335
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.11
# d = 141
# alternative fragment length(s) may be 141 bps
```





chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)	name
chr1	2603	2989	387	2870	18.00	6.68596	3.52825	3.66748	AR_peak_1
chr1	138179	138371	193	138281	18.00	14.90779	7.93021	11.47829	AR_peak_2
chr1	36515	36714	200	36609	16.00	12.59143	7.05394	9.25447	AR_peak_3
chr1	201091	201231	141	201114	10.00	7.58293	5.23859	4.50002	AR_peak_4
chr1	69373	69558	186	69452	18.00	9.61904	4.93737	6.41821	AR_peak_5



MACS: Output interpretation

- Excel

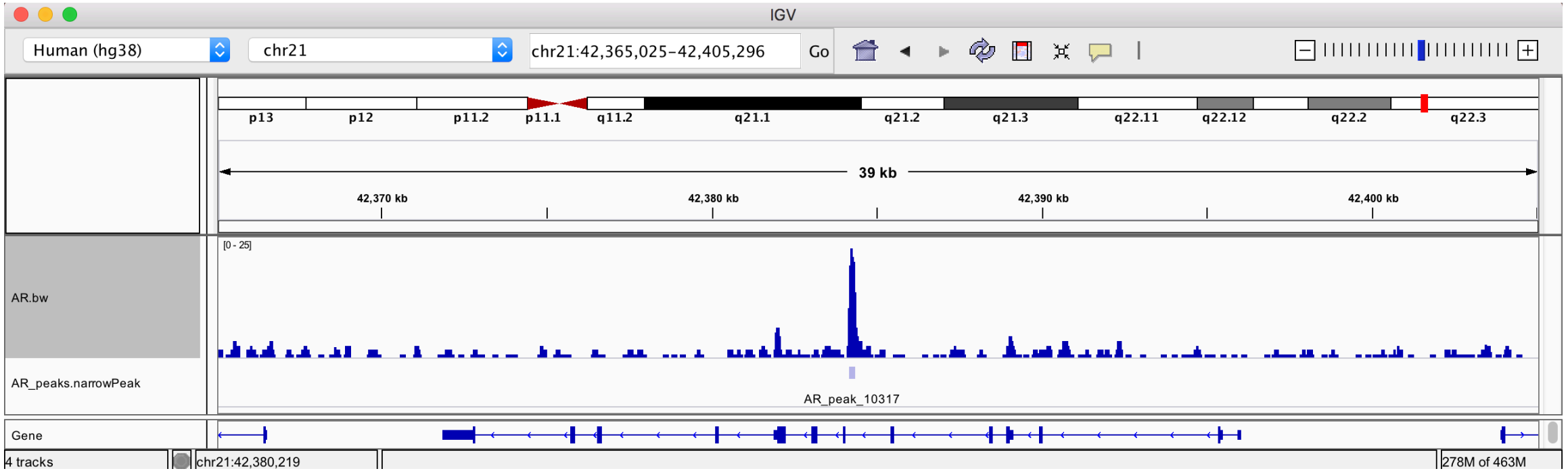
chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-log10(qvalue)	name
chr1	2603	2989	387	2870	18.00	6.68596	3.52825	3.66748	AR_peak_1
chr1	138179	138371	193	138281	18.00	14.90779	7.93021	11.47829	AR_peak_2
chr1	36515	36714	200	36609	16.00	12.59143	7.05394	9.25447	AR_peak_3
chr1	201091	201231	141	201114	10.00	7.58293	5.23859	4.50002	AR_peak_4
chr1	69373	69558	186	69452	18.00	9.61904	4.93737	6.41821	AR_peak_5

- narrowPeak

				score		fold	p	q	sm
chr1	591170	591325	AR_peak_290	82	.	6.63900	11.50806	8.21785	25
chr1	629218	629993	AR_peak_291	295	.	3.42374	33.50185	29.54851	636
chr1	630286	630453	AR_peak_292	106	.	2.39458	14.04047	10.64496	81
chr1	630765	631382	AR_peak_293	239	.	3.14283	27.79379	23.97848	480
chr1	631877	632366	AR_peak_294	224	.	3.06645	26.24850	22.47273	380

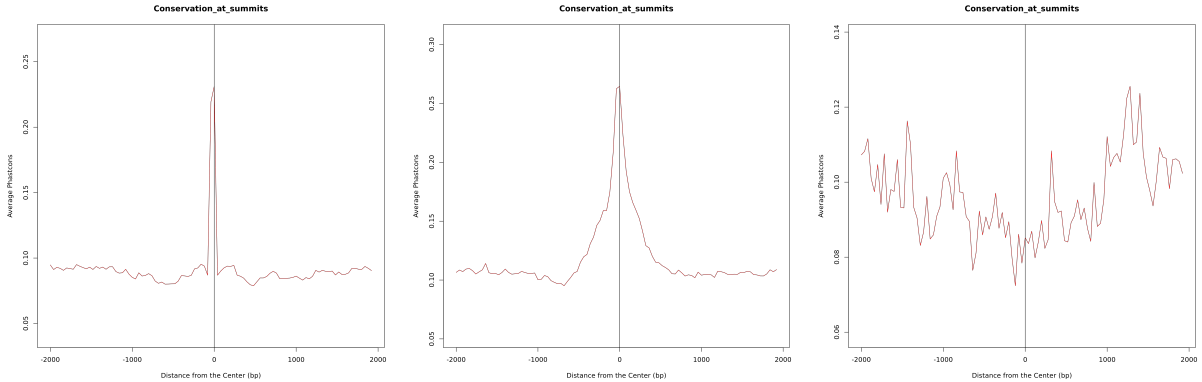
Data Visualization

- bedGraph to bigWig
- macs2 output data
- IGV

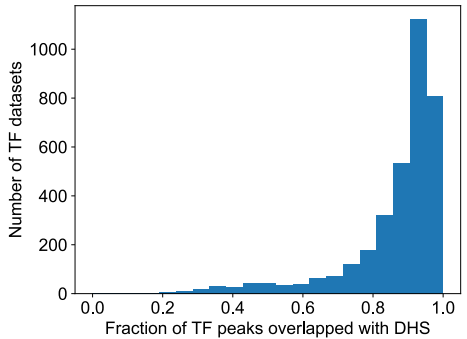


Quality Control

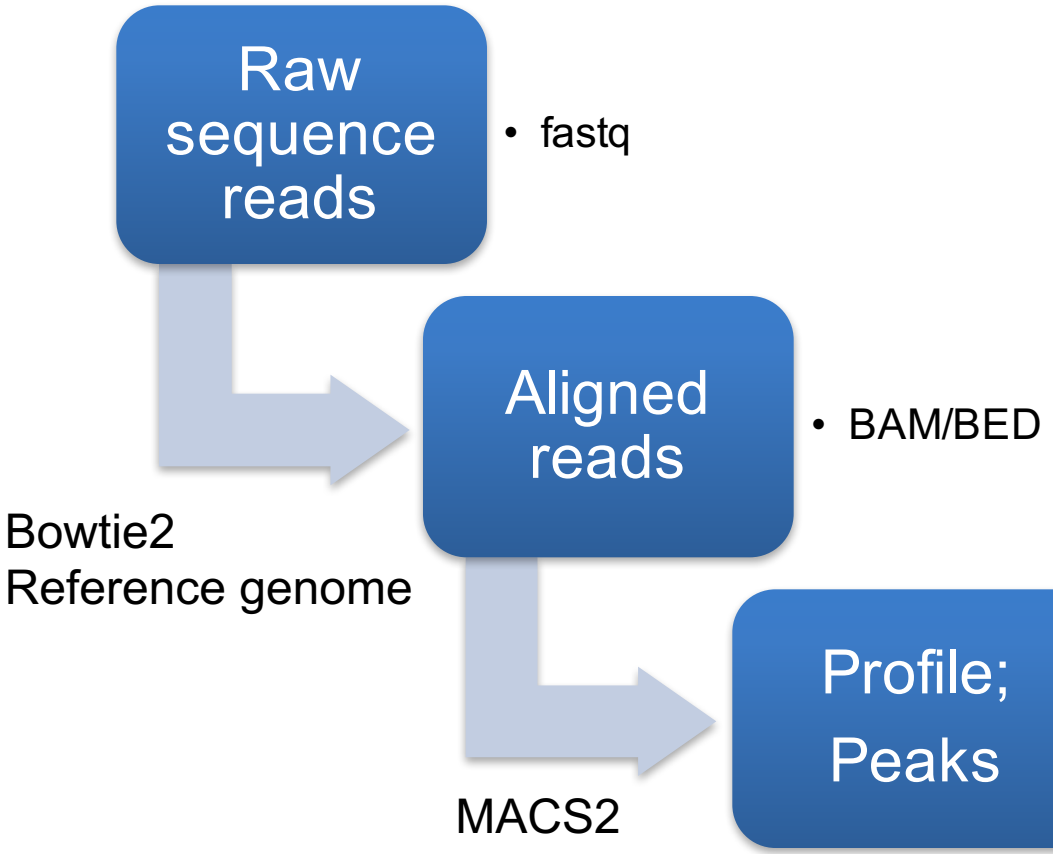
- FRiP (Fraction of Reads in Peaks) score
  - 1-10% for TF is normal
- Number of peaks
  - Number of peaks with high fold-enrichment, e.g, 5, 10, …
  - 2000
- Sequence conservation



- Fraction of peaks within regulatory regions
  - 80%

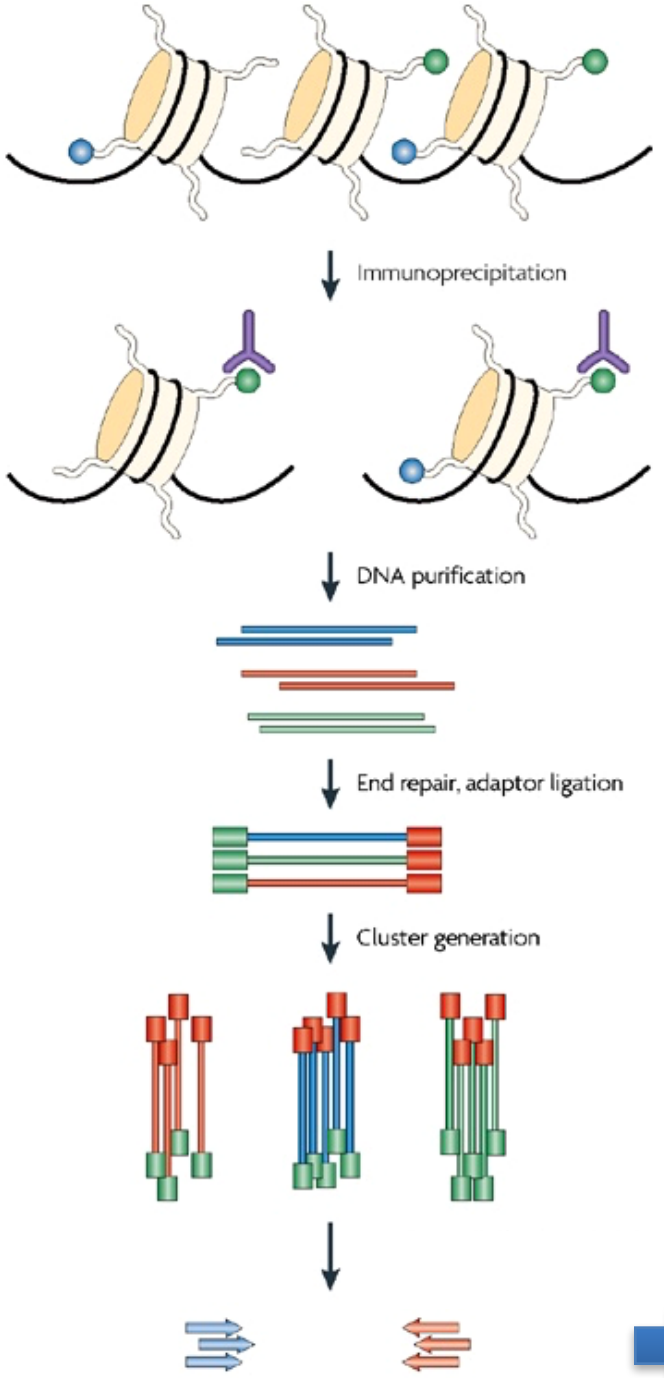


Data flow and QC summary



QC measures

- Sequence quality (fastqc)
- Mapping quality (uniquely mapped ratio)
- Library complexity (PBC)
- Fold enrichment, peaks
- Signal/Noise (FRIP score)
- Regulatory annotation



Experimental procedure

Biology

downstream analysis/integration

Peaks (bed)

peak calling

Pile-up for visualization
(bedGraph, wig, bigwig)

macs2

Non-redundant reads (sam/bam/bed)

redundancy assessment

Mapped reads (sam/bam/bed)

alignment (**bowtie2**)

Raw sequence reads (fastq)

Computational analysis

Questions?

zang@virginia.edu
zanglab.org