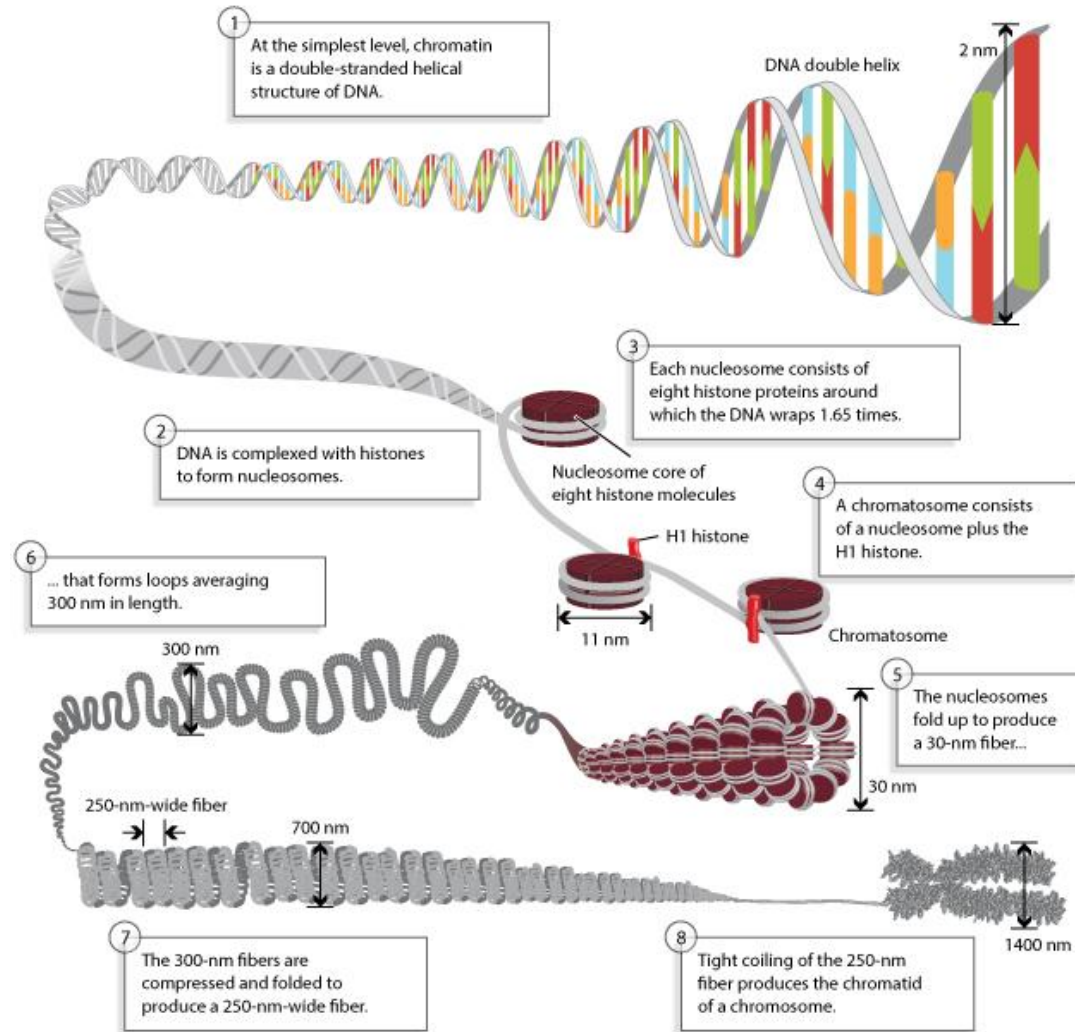


Introduction to ChIP-seq

Overview, alignment, peaks, and motifs

DNA hierarchy

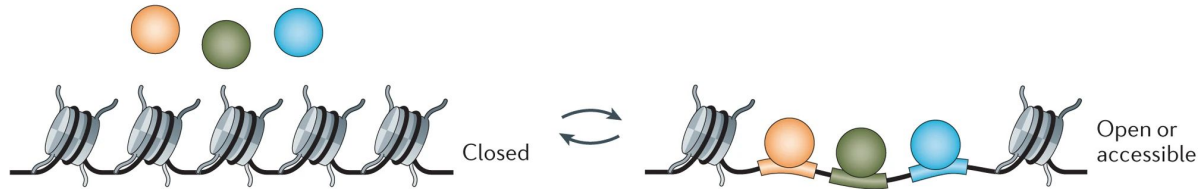
- Multiple levels of organization for DNA, histones, and chromatin
- For ChIP-seq, our interest is in levels 2 and 3
- Related assays probe for chromatin accessibility



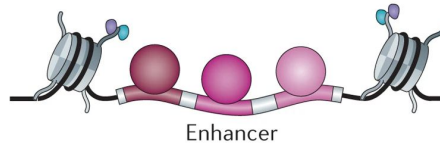
Gene regulation

- Chromatin elements play a vital role in regulating gene expression

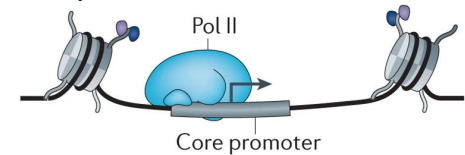
a Chromatin as accessibility barrier



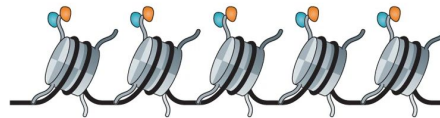
b Active enhancer



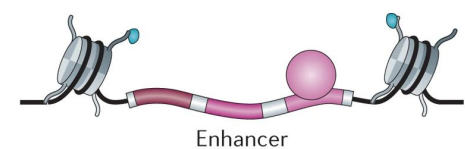
c Active promoter



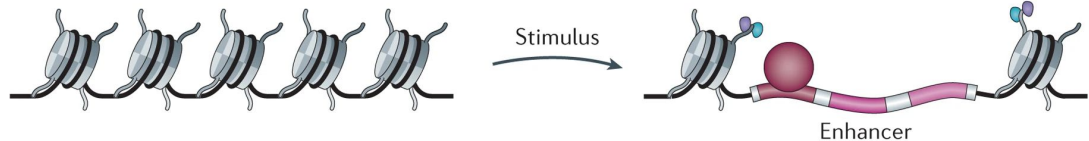
d Closed or poised enhancer



e Primed enhancer



f Latent enhancer



DNA binding motifs



DNA-binding proteins:
TFs, CTCF, repressors
and polymerases

H3K4me1
H3K4me3

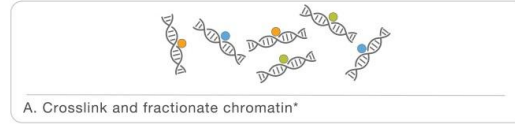
H3K27ac
H3K27me3

ChIP-seq protocol

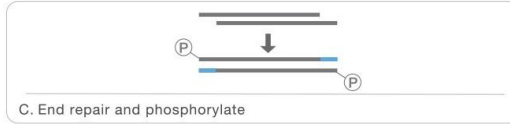
- An assay for studying gene regulation
 - Promoters
 - Enhancers
 - Repressors
- Profiling protein binding (ChIP) of DNA (seq)
 - Transcription factors
 - Histone modifiers



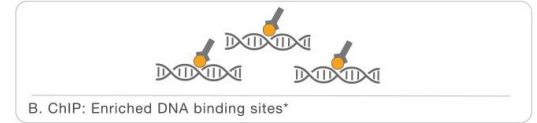
Nucleus



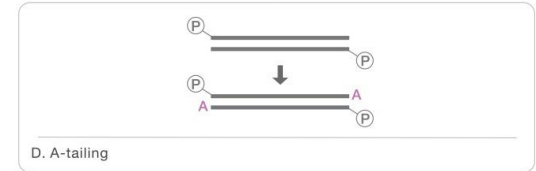
A. Crosslink and fractionate chromatin*



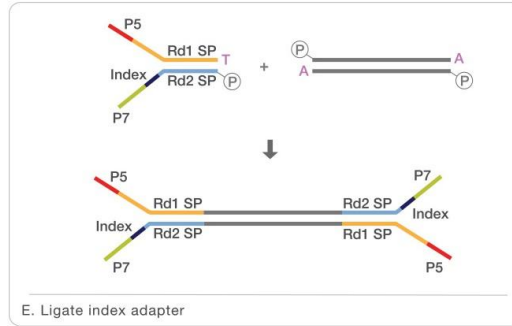
C. End repair and phosphorylate



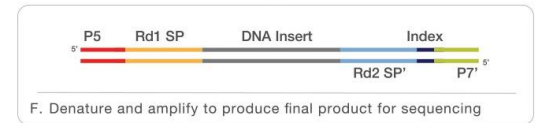
B. ChIP: Enriched DNA binding sites*



D. A-tailing



E. Ligate index adapter

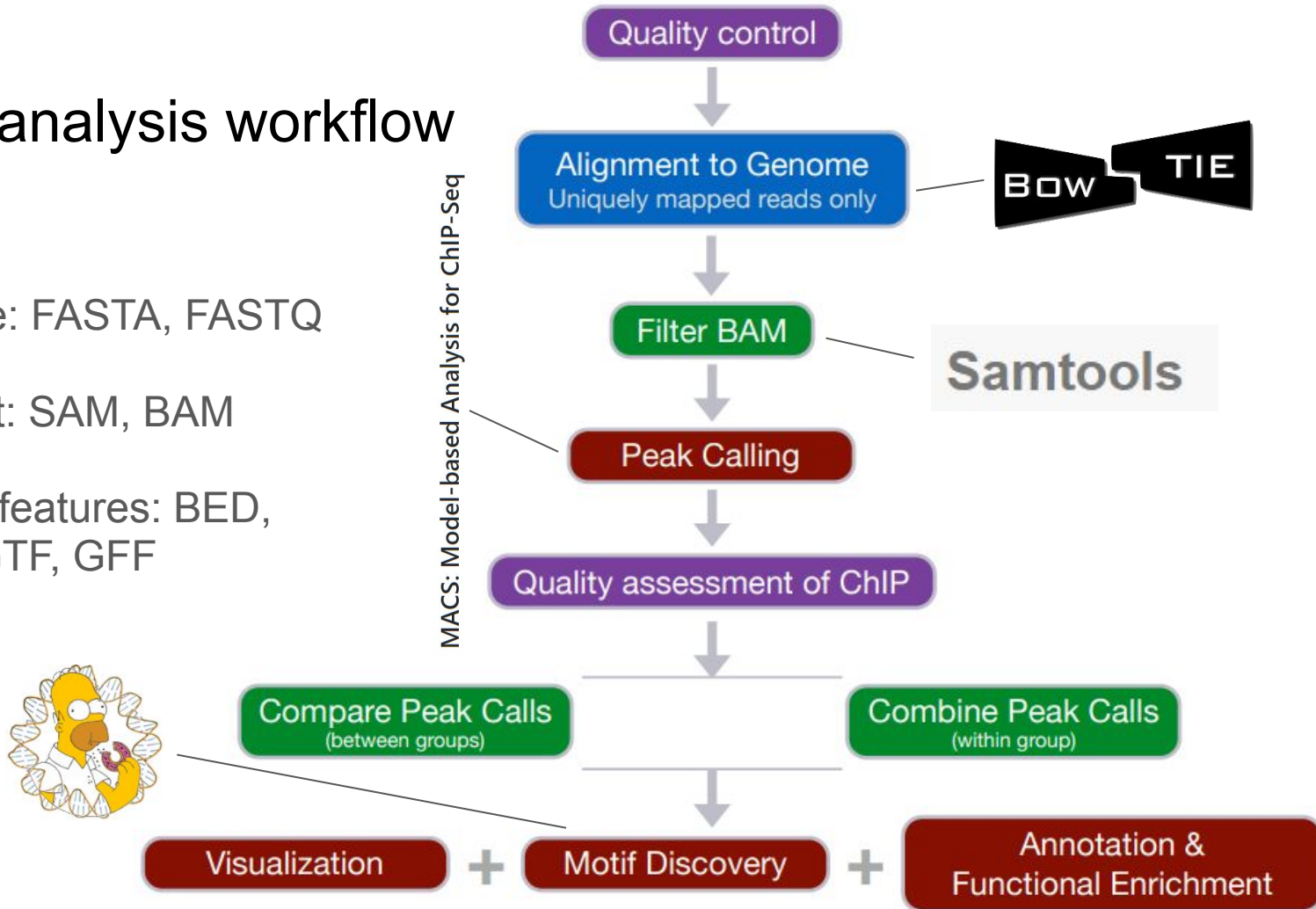


F. Denature and amplify to produce final product for sequencing

ChIP-seq analysis workflow

Data types

1. Sequence: FASTA, FASTQ
2. Alignment: SAM, BAM
3. Genomic features: BED, Wiggle, GTF, GFF



BED file format

1. chrom: name of the chromosome
2. chromStart: starting position of the feature (0-base)
3. chromEnd: ending position of the feature

(optional columns 4-6)

4. name: name for the BED line
5. score: quality score (sometimes omitted in place of something else)
6. strand: “+” (sense), “-” (antisense), “.” (NA)

(optional columns 7-12, see [UCSC Genomics Institute page](#))

chr1	213941196	213942363			
chr1	213942363	213943530			
chr1	213943530	213944697			

chr7	127471196	127472363	Pos1	0	+
chr7	127472363	127473530	Pos2	0	+
chr7	127473530	127474697	Pos3	0	+

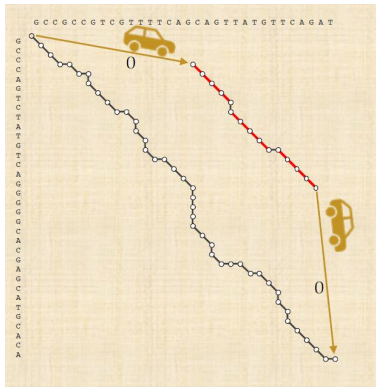
Chromosome ID →

Start location ↑ End location ↑ Name ↑ Strand ↑

Phase (reading frame)

Read alignment

- Bowtie2, Bowtie1, and bwa are all suitable methods dependent on read length
- We have to construct an index
- Can perform local or global alignment



Local Alignment

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||
5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||
5' ACTACTAGATT-----ACGGATC---GTACTTTAGAGGCTAGCAACCA 3'
```

Quality control

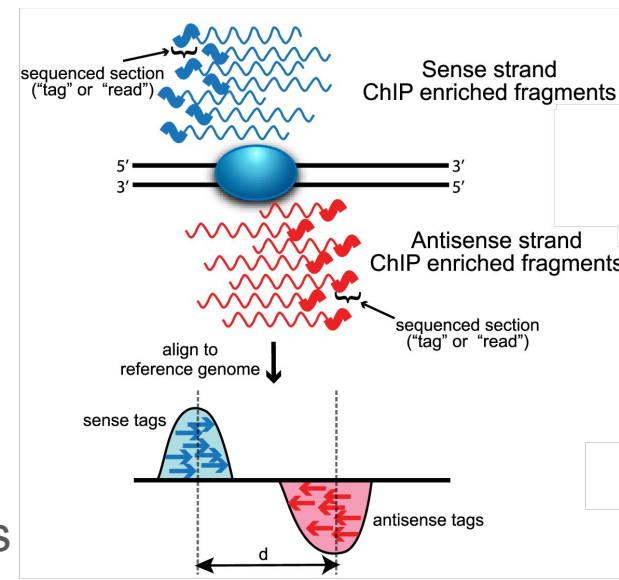
```
[zemke@n7420 flashscratch]$ bowtie2 -q -p 4 -k 1 --local  
mso_k18_chip.fastq > dmso_k18_chip.sam  
42714660 reads; of these:  
  42714660 (100.00%) were unpaired; of these:  
    980458 (2.30%) aligned 0 times  
    41734202 (97.70%) aligned exactly 1 time  
    0 (0.00%) aligned >1 times  
97.70% overall alignment rate
```

- Non-redundant fraction
 - How many reads were uniquely mapped?
- Fraction of reads in peaks
 - Proportion of reads in significant regions of the genome?
 - Reference peak set
- Blacklist region
 - Regions of the genome where reads should be thrown out

The ENCODE Blacklist: Identification of Problematic Regions of the Genome

Peak calling

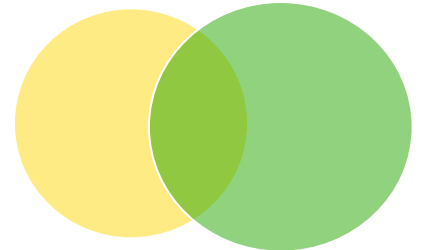
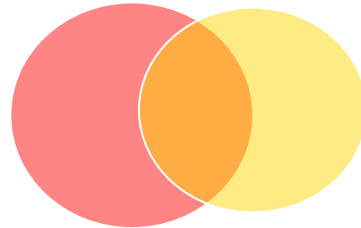
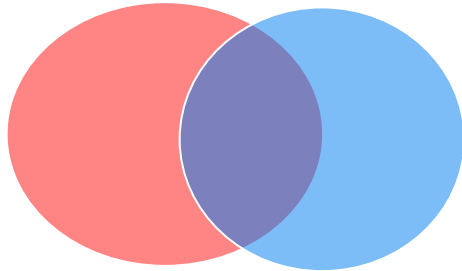
- We want to identify regions of the genome where many aligned reads are found
- MACS tests whether the background is Different from peaks using Poisson models
- Duplicates should be removed prior to peak calling otherwise there will be false positives



Differential peaks

Peaks that are unique to some set may be associated with specific motifs

	H3K27ac (enhancer)	H3K4me3 (repressor)
treatment	A / input	C / input
control	B / input	D / input



Motif analysis

- Peaks that are differentially expressed may have specific sequence motifs
- HOMER compares the peak set against a background set or reference genome using the binomial distribution
 - Consider the search range around peak for identifying motifs