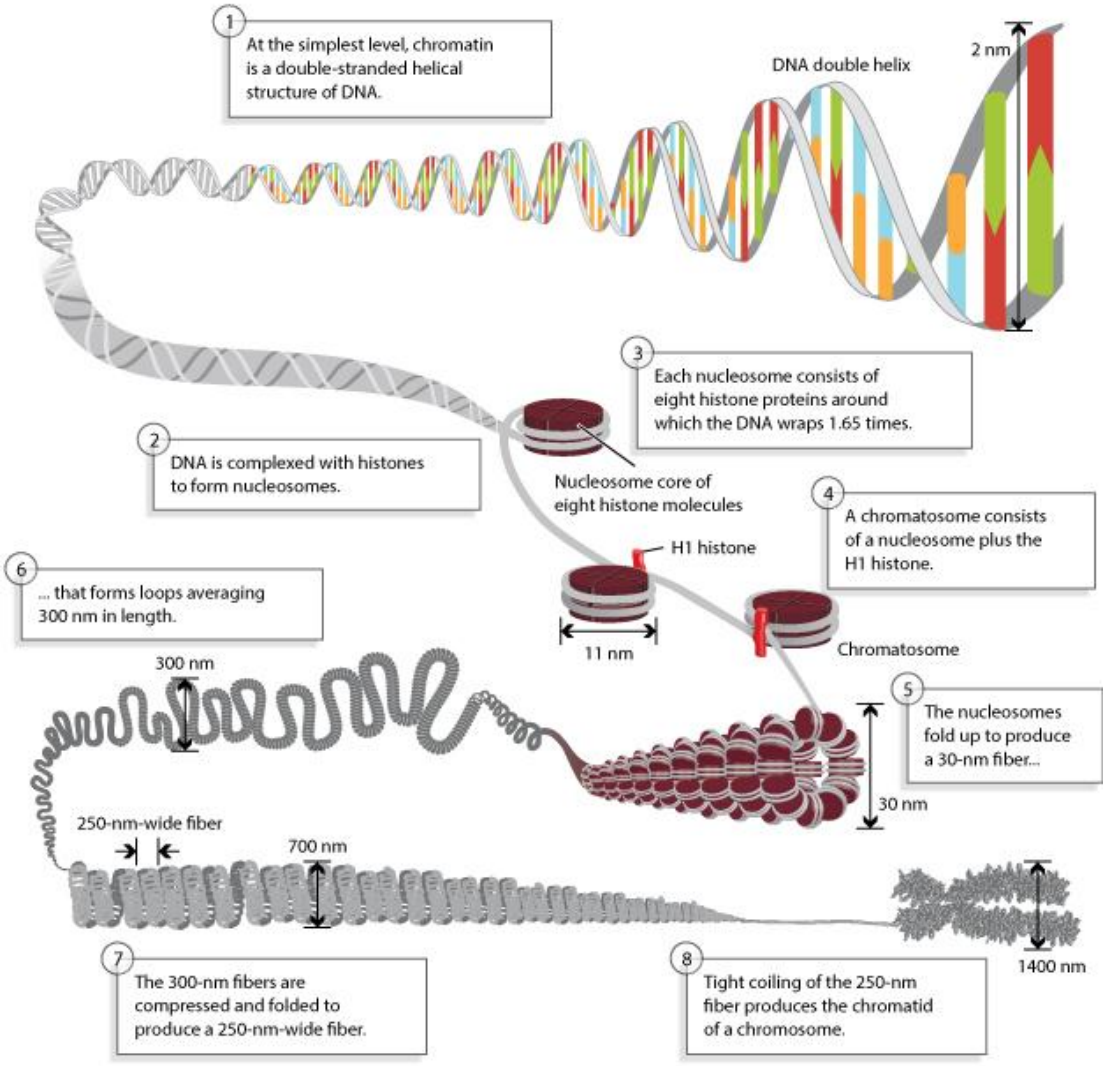
Introduction to ChIP-seq

Overview, alignment, peaks, and motifs

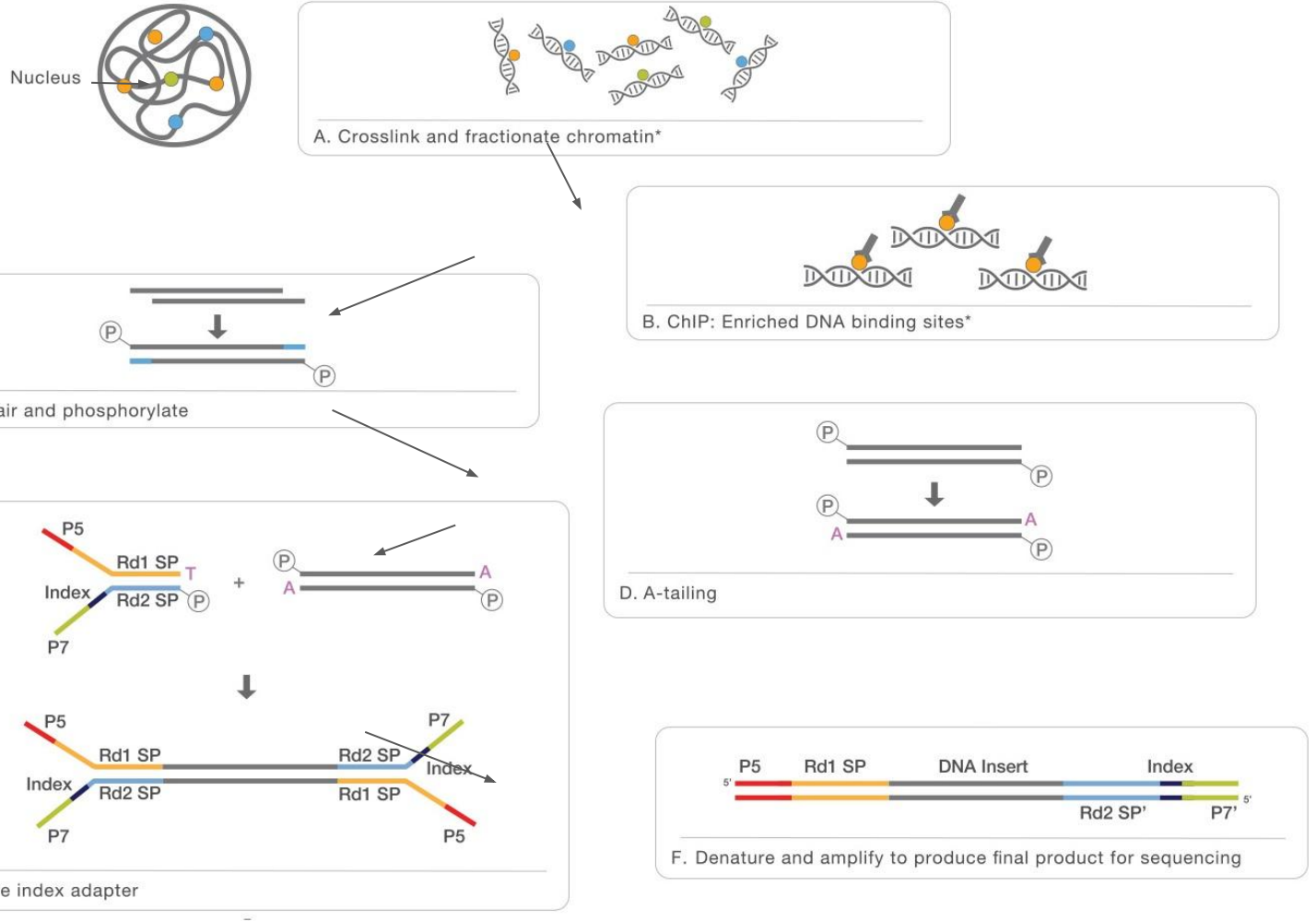
What is ChIP-seq?

- A tool for studying gene regulation: promoters, enhancers, and repressors
- Assay profiling protein binding (chromatin immunoprecipitation) to DNA (sequencing)
- Transcription factor binding and histone modifier interactions to DNA

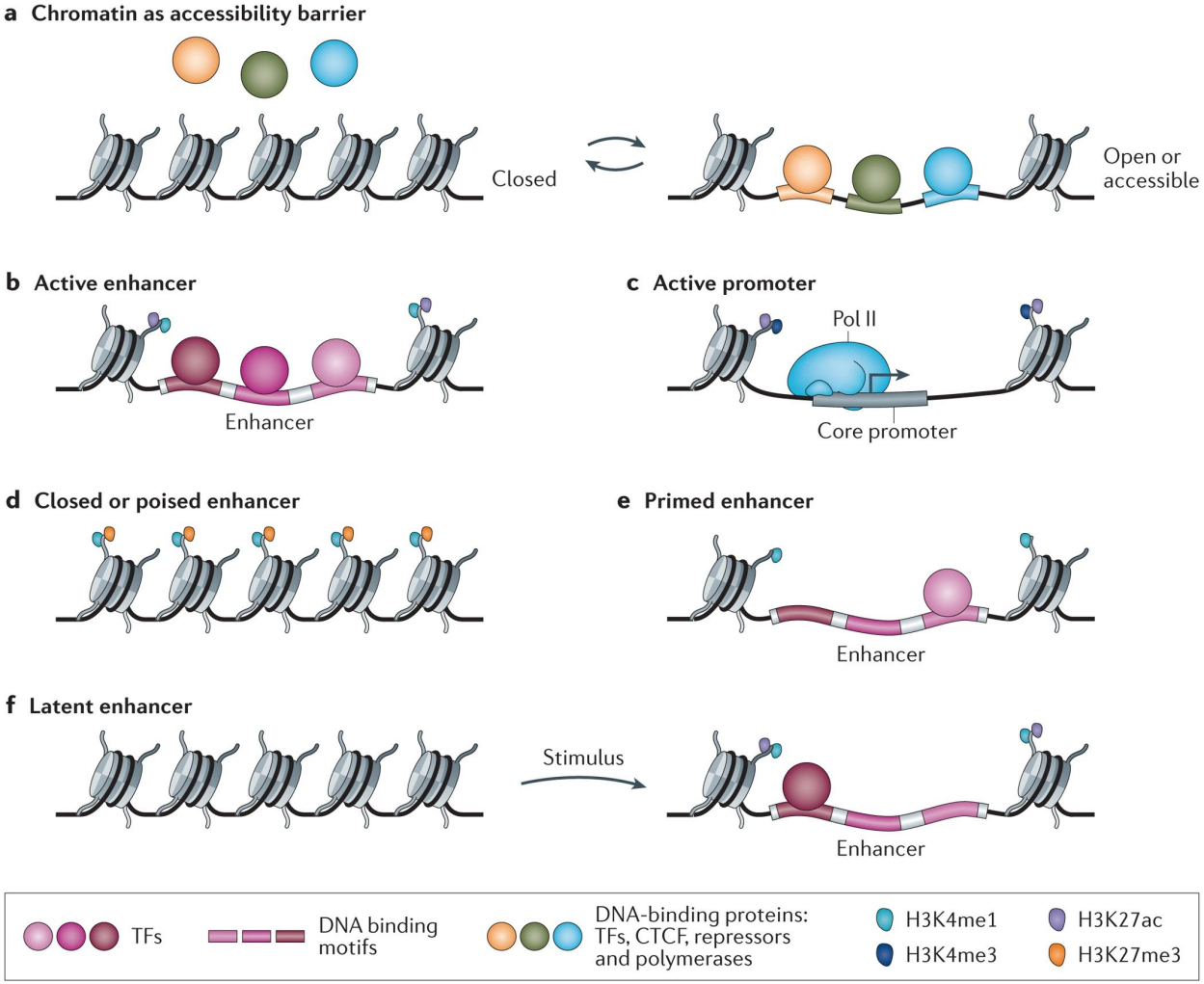
DNA hierarchy



ChIP-seq protocol



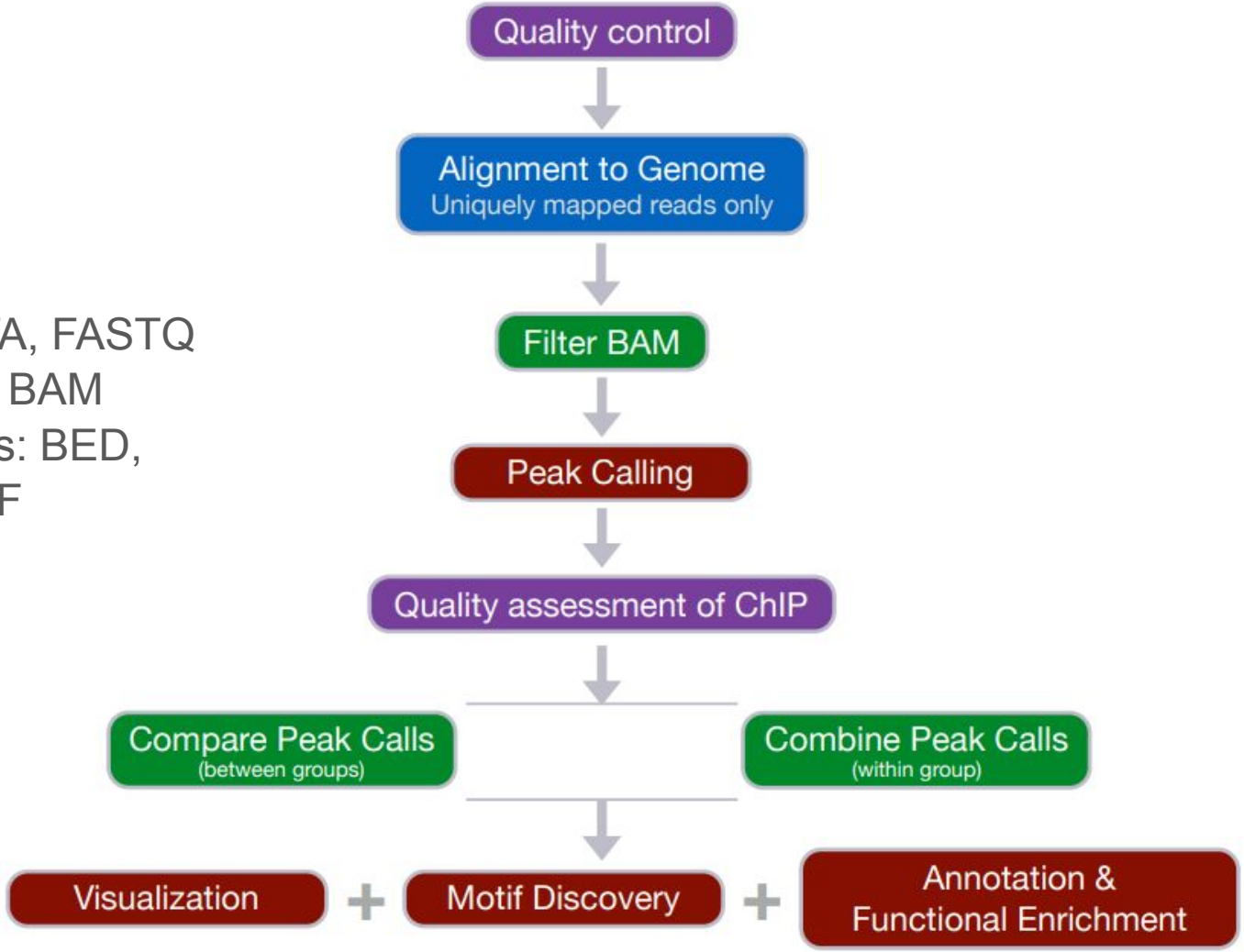
Gene regulation



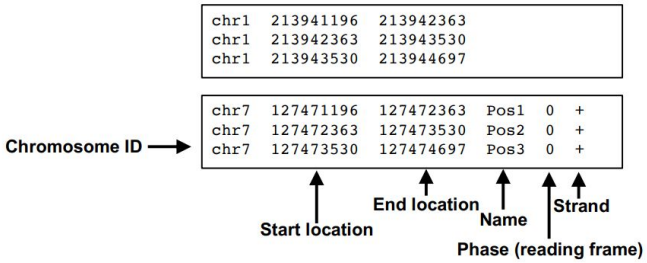
Workflow

Data types

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



BED file format



1. Chr: name of the chromosome
2. Start: starting position of the feature (0-base)
3. End: 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)

Alignment

Local Alignment

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

Global Alignment

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

- Bowtie2 and bwa are both suitable methods
- Like before, we have to construct an index
- Can perform local or global alignment

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
```

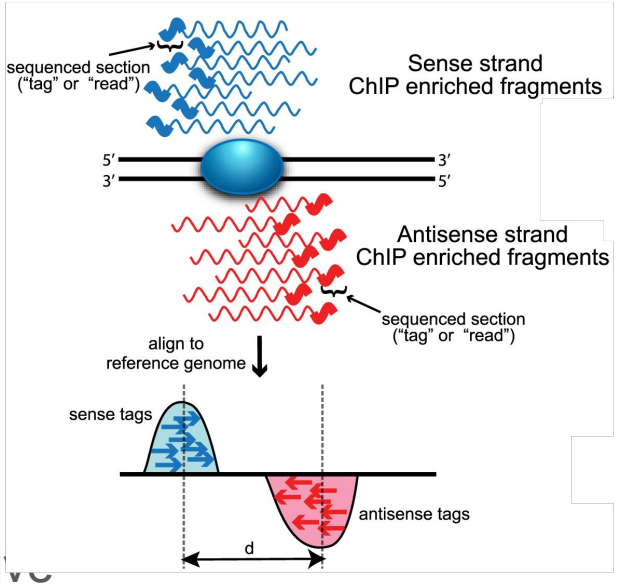
Non-redundant fraction (NRF)

- How many reads were mapped uniquely?

Fraction of reads in peaks (FRiP)

- How many reads were in significant regions of the genome? (can use a reference set)

Peak calling

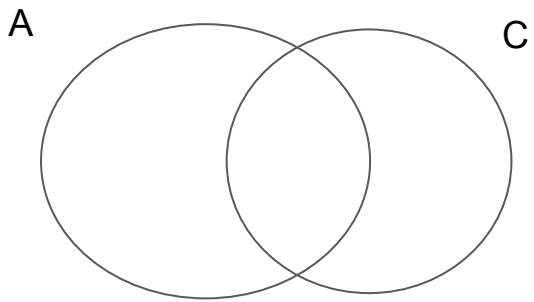


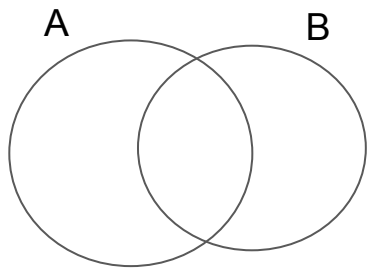
- We want to identify regions of the genome where many aligned reads are located
- MACS2 fits Poisson models to the background and tests if some prospective peak region follows a separate Poisson model
- Duplicates need to be removed prior to peak calling otherwise false signals will be identified

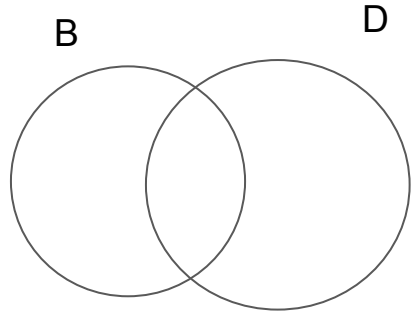
Differential peaks

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 enrichment

HOMER

- Can compare peak set with a background peak set or use a reference genome
- Significantly enriched if fraction of sequences in input matching motif is significantly different from fraction of sequences in background

Additional resources

https://genome.ucsc.edu/FAQ/FAQformat.html