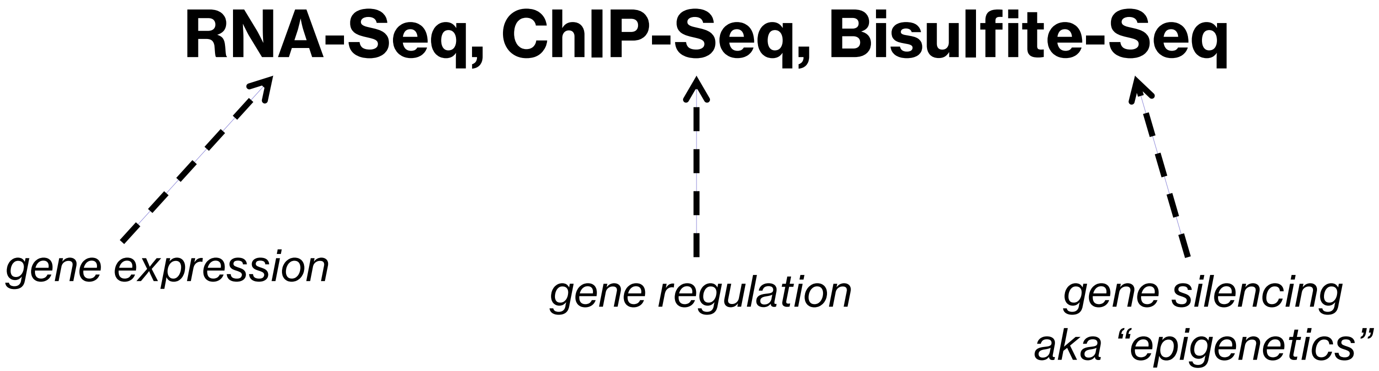
Biochem 3BP3

RNA-Seq, ChIP-Seq, Bisulfite-Seq

Week of Nov 15, 2021

Biochem 3BP3



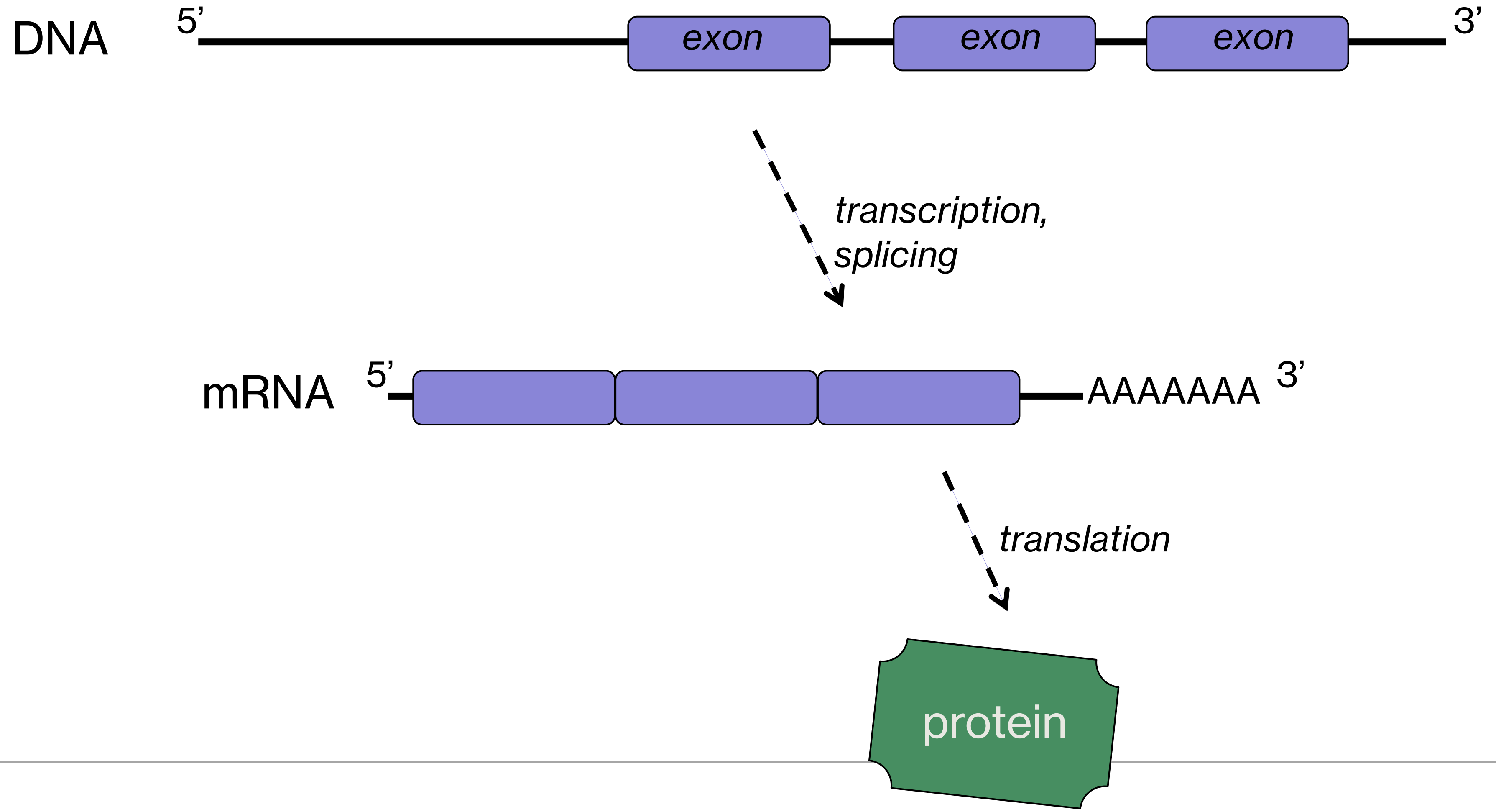
RNA-Seq, ChIP-Seq, Bisulfite-Seq

gene expression

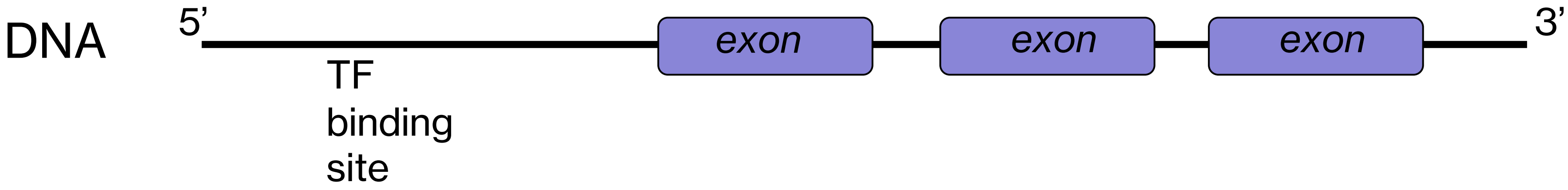
gene regulation

gene silencing
aka “epigenetics”

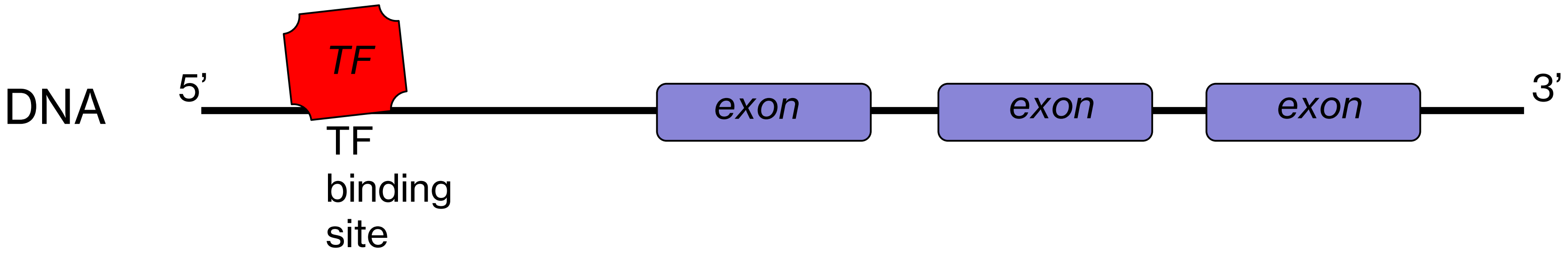
Gene Expression



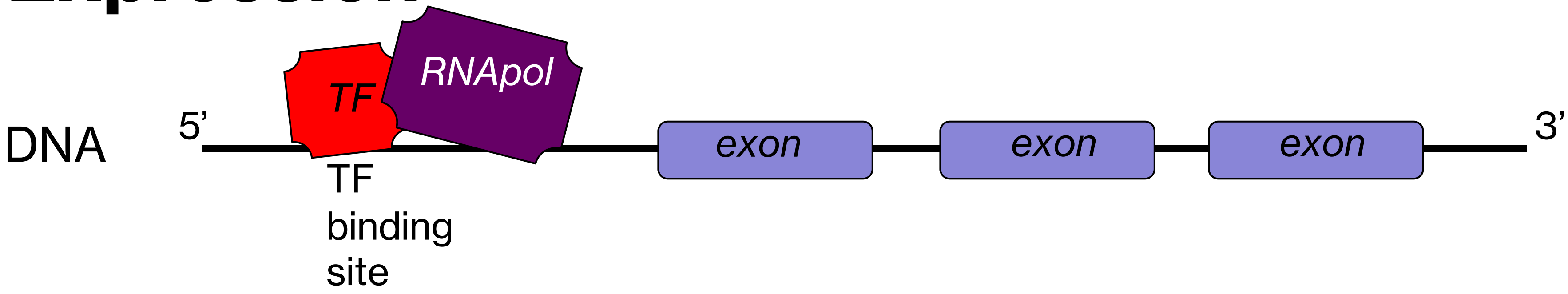
Gene Expression



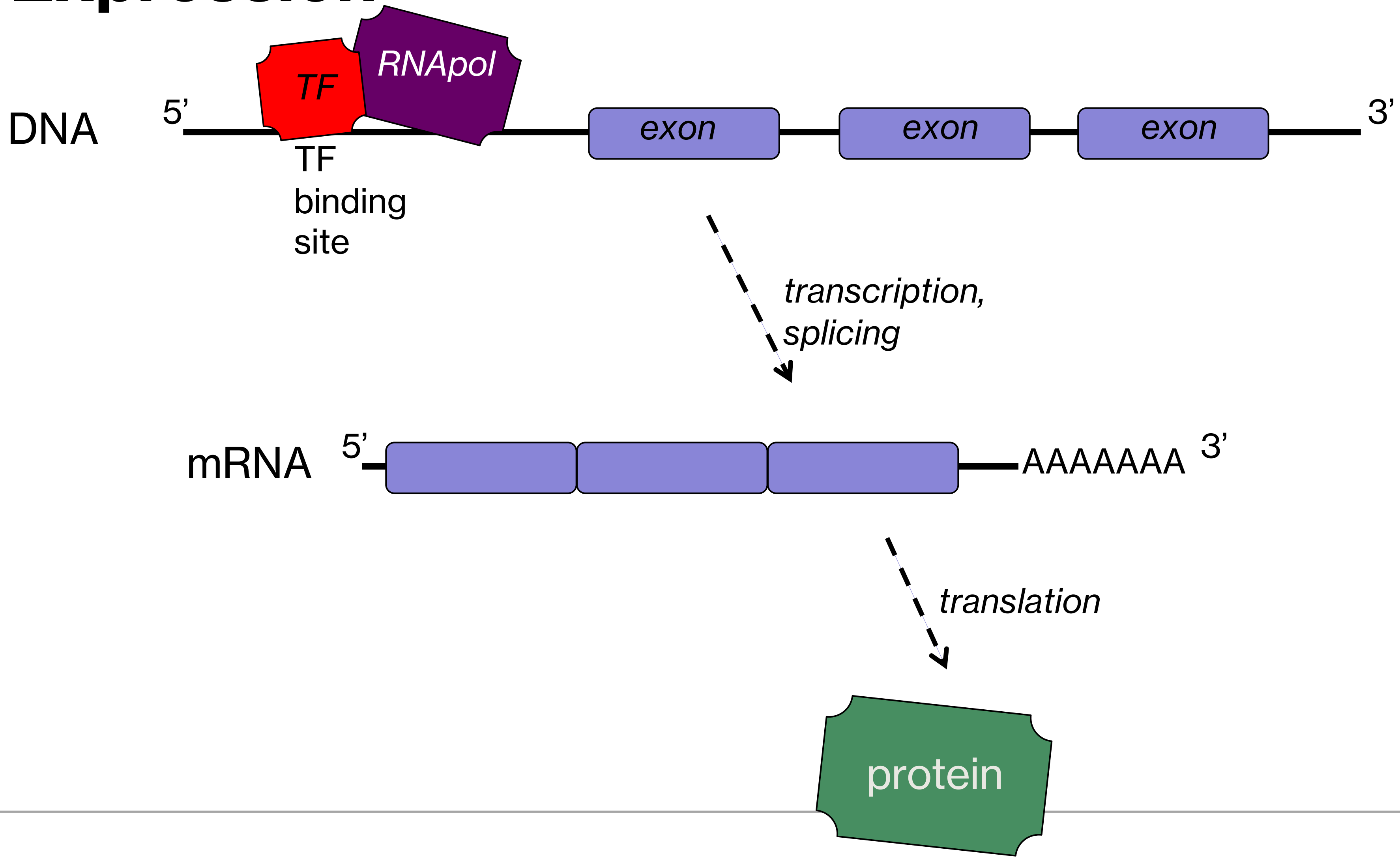
Gene Expression



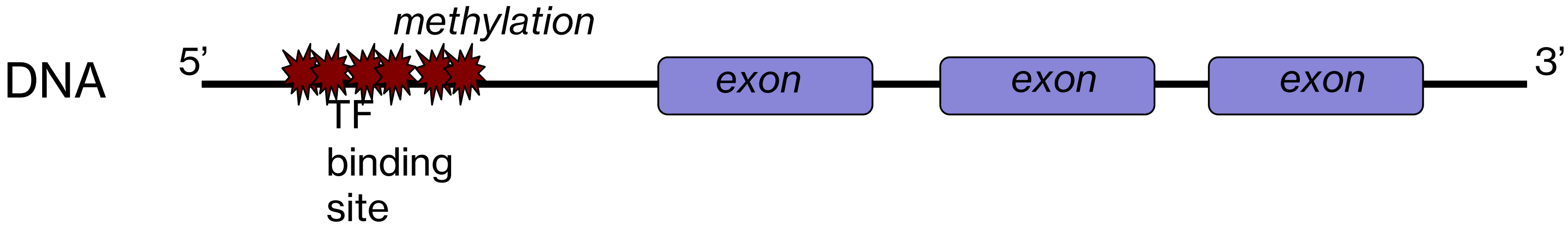
Gene Expression



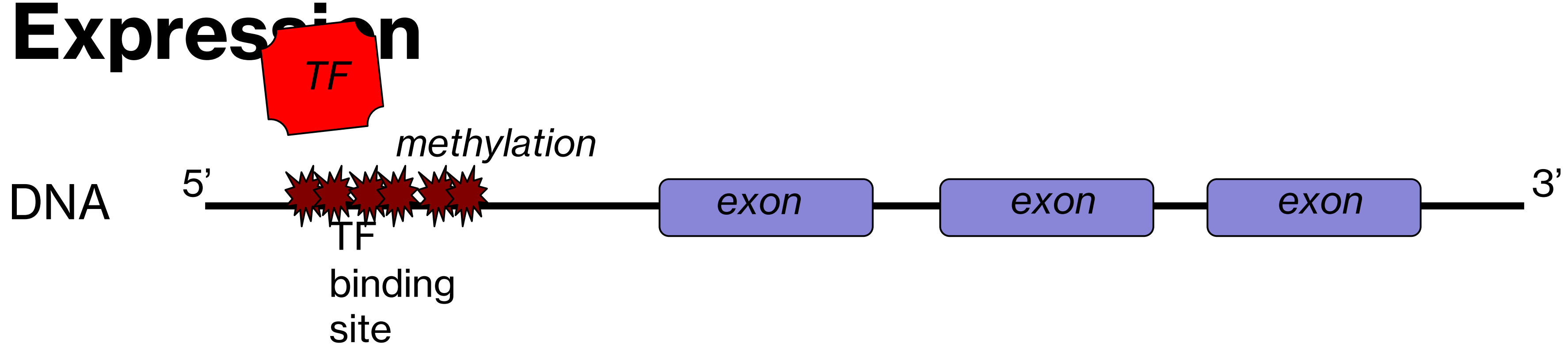
Gene Expression



Gene Expression



Gene Expression



gene silencing via genomic DNA methylation

Transcription, Regulation, Epigenetics

- RNA-Seq is a NGS method to estimate the abundance of mRNA transcripts in the transcriptome

- ChIP-Seq is a NGS method to detect transcription factor or other DNA-binding protein binding sites in the genome

- Bisulfite-Seq is a NGS method to determine methylation patterns in the genome

Transcription, Regulation, Epigenetics

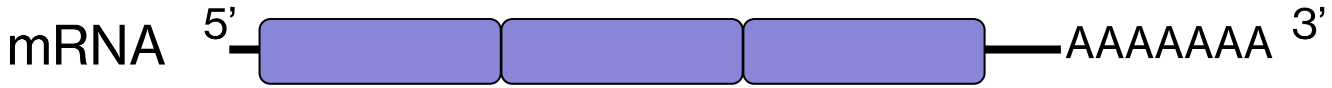
Each method relies on the generation of millions of NGS reads

- These are DIRECT methods!
- How the reads are generated determines what is being measured – novel library preparation methods
- All three methods rely heavily on the alignment of short NGS reads to reference genome sequences using variants of the Burrows-Wheeler Transform (e.g. BWA, Tophat, HiSAT2, etc.)
- Depth of sampling of NGS reads will vary among replicates and the data thus require normalization
- Some of the methods rely heavily upon use of technical controls (i.e. a control for the methodology, not the experimental conditions)

What is a technical control?

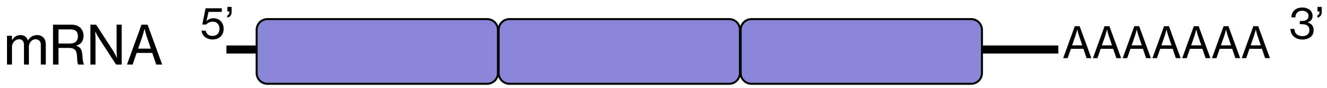
- A **technical control** corrects for steps in the **assay** being used. It removes background noise or bias in the technology, resulting in corrected values being used for statistical tests.
- In the microarray lab, we made technical corrections without explicit use of technical controls:
  - Background subtraction in microarrays to remove background fluorescence of glass
  - Non-linear normalization across microarrays due to loading error
- An **experimental control** corrects for steps in the **experiment** to focus the observed values on the effect being studied. For example:
  - Zebrafish embryos exposed to TCDD (dissolved in DMSO)
  - Zebrafish embryos exposed to DMSO only

RNA-Seq



- Isolate mRNA using the same methods as for microarray experiments
- Fragment the mRNA sample into small pieces of mRNA (or cDNA), link to Illumina adapators, and perform high-throughput DNA sequencing
- Non-biased sampling of the transcriptome
  - mRNA fragments are randomly distributed across transcripts
  - abundance of fragments is reflective of actual transcript abundance

RNA-Seq



As in genome sequencing, transcriptome sequencing assembly is improved by using Illumina mate-pairs (i.e. forward & reverse reads)

Unlike genome sequencing, length of NGS reads is less important

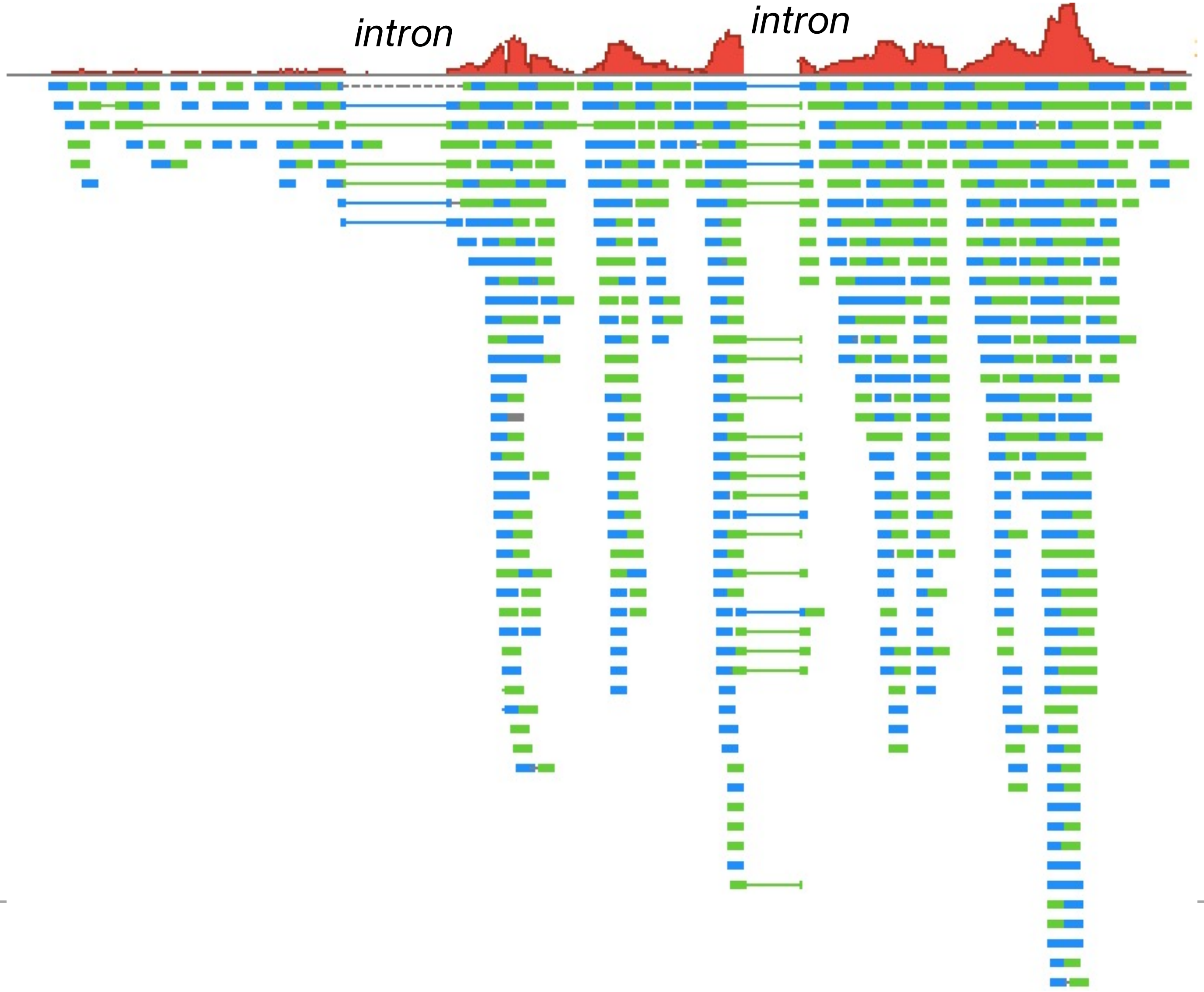
- using small fragment size libraries
- assembly challenge is local (i.e. a transcript) instead of global (i.e. a genome)
- depth of sampling is **critical** for accurate assessment of transcriptome abundance and there is more to sample than whole-genome shotgun
- often see 2 x 50 bp Illumina sequencing = cost less = more depth

RNA-Seq

- NGS reads are mapped to the genome sequence using mRNA specific variants (e.g. HiSAT2) of the Burrows-Wheeler Transform since RNA-Seq reads will not contain intron sequences

- RNA-Seq thus not only measures transcript abundance but it also helps define transcript models – intron/exon boundaries – as well as detect alternate splicing

RNA-Seq



RNA-Seq

- Since each RNA-Seq mate pair is generated from a single mRNA fragment, a mate-pair is the base unit for measurement of transcript abundance
- Linear normalization methods adjust for sampling effort and gene size; FPKM = fragments per kilobase of transcript per million mapped reads; for example:
    - 11,000,000 RNA-Seq reads generated
    - 8,300,000 RNA-Seq reads mapped to genome
    - 783 fragments mapped to CYP19
    - CYP19 transcript is 913 bp
    - FPKM = 783 / (913/1000) / (8300000 / 1000000) = 103.32
- As observed in microarray experiments, abundance plots between samples are often non-linear, requiring more sophisticated normalization (next slide)
- Final normalized data sets are submitted to standard statistical tests very similar to single-dye microarray data
    - avoid use of fold-change statistics
    - considerable focus on multiple tests and false discovery rates

RNA-Seq Normalization

- Single channel microarrays – Cy3 intensity converted to a number by Feature Extraction: bigger number = more mRNA
- RNA-Seq – mRNA fragments (i.e. mate pairs) aligned to genes and # fragments per gene counted: higher count = more mRNA
- In microarrays, we had to normalize for loading error: different amounts of Cy3 labeled cDNA pipetted onto each microarray
- In RNA-Seq, after removal of low quality sequence based on PHRED values plus exclusion of contaminant sequences that do not map to the reference genome, each library ends up sequenced to slightly different depths: **this is equivalent to loading error**
- Normalization via Rank Invariant Probes would work for both single channel microarrays and RNA-Seq data
- However, RNA-Seq normalization algorithms have advanced further…

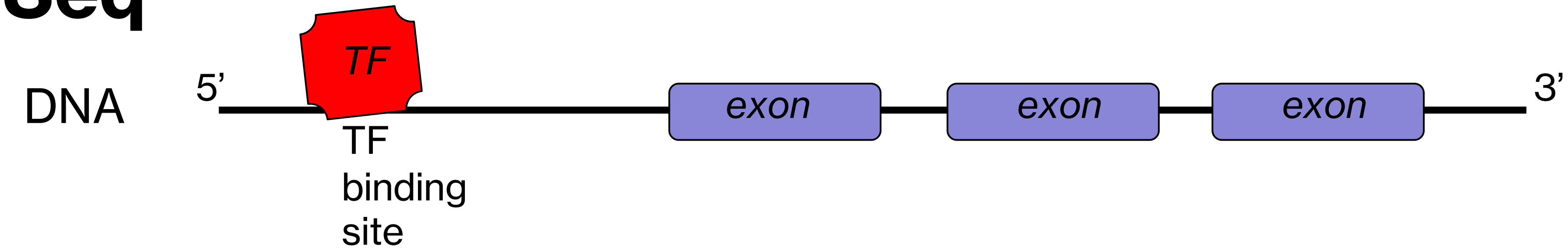
DESeq2 for RNA-Seq

- DESeq2 performs non-linear normalization of RNA-Seq data using the negative binomial distribution & local regression
- Normalization algorithm down-plays highly expressed or highly variable genes and thus reduces Type I Error (false discovery), particularly for lowly or highly expressed genes
- DESeq2 can perform both normalization and significance tests
- DESeq2 significance testing has some of the best additional FDR correction algorithms for RNA-Seq as well as use of General Linear Models to handle complex experimental designs
    - Pre-test exclusion of non-variable genes to reduce the number of tests performed
    - Benjamini and Hochberg correction for FDR for the remaining genes

DESeq2 for RNA-Seq

- DESeq2 performs non-linear normalization of RNA-Seq data using the negative binomial distribution & local regression
- Normalization algorithm down-plays highly expressed or highly variable genes and thus reduces Type I Error (false discovery), particularly for lowly or highly expressed genes
- DESeq2 can perform both normalization and significance tests
- DESeq2 significance testing has some of the best additional FDR correction algorithms for RNA-Seq as well as use of General Linear Models to handle complex experimental designs
    - Pre-test exclusion of non-variable genes to reduce the number of tests performed
    - Benjamini and Hochberg correction for FDR for the remaining genes

ChIP-Seq

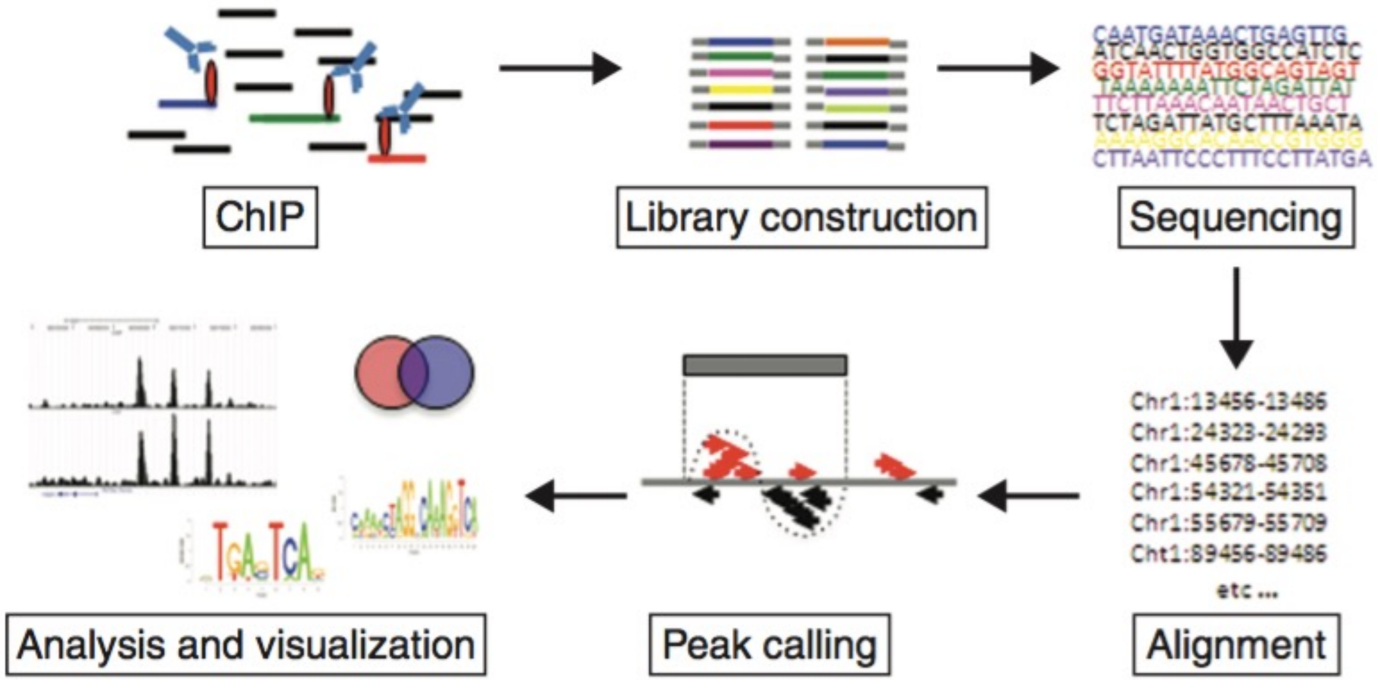


- ChIP-Seq = chromatin immunoprecipitation (ChIP) assays with NGS

- Goal is to determine locations of DNA binding for a specific protein

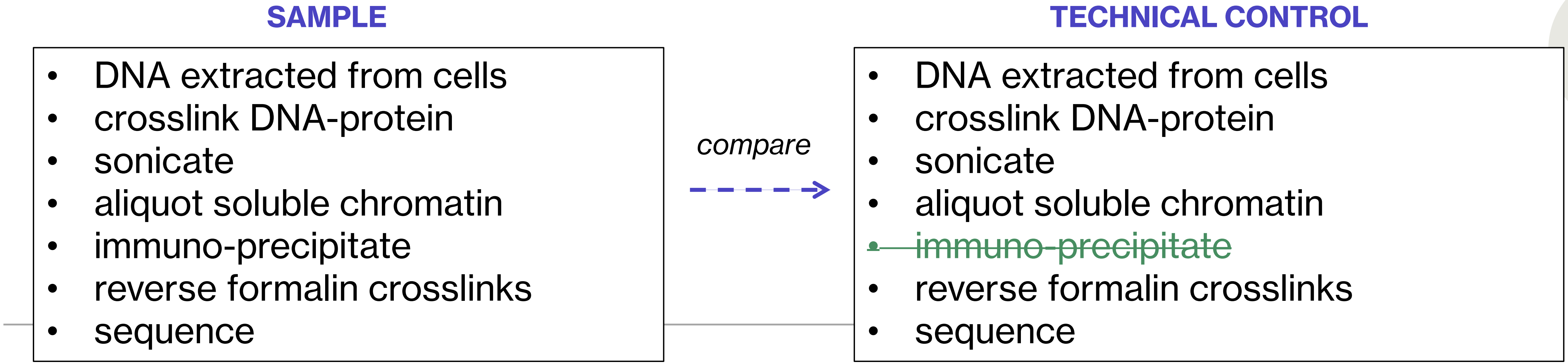
ChIP-Seq

- DNA extracted from cells – some of the DNA will have proteins attached
- formalin crosslink DNA-protein so proteins cannot fall off
- sonicate to destroy DNA not cross-linked to a protein
- aliquot soluble chromatin
- immuno-precipitate using an antibody specific to target protein
- reverse formalin crosslinks
- sequence



ChIP-Seq

- NGS reads acquired from ChIP-Seq should only be from the locations where the target protein was bound to the genome
- The quality and specificity of the antibody is paramount!
  - ChIP-Seq data is often noisy due to low rates of non-specificity
- Background noise from incomplete sonication of non-bound DNA is corrected by use of a technical control



SAMPLE

- DNA extracted from cells
- crosslink DNA-protein
- sonicate
- aliquot soluble chromatin
- immuno-precipitate
- reverse formalin crosslinks
- sequence

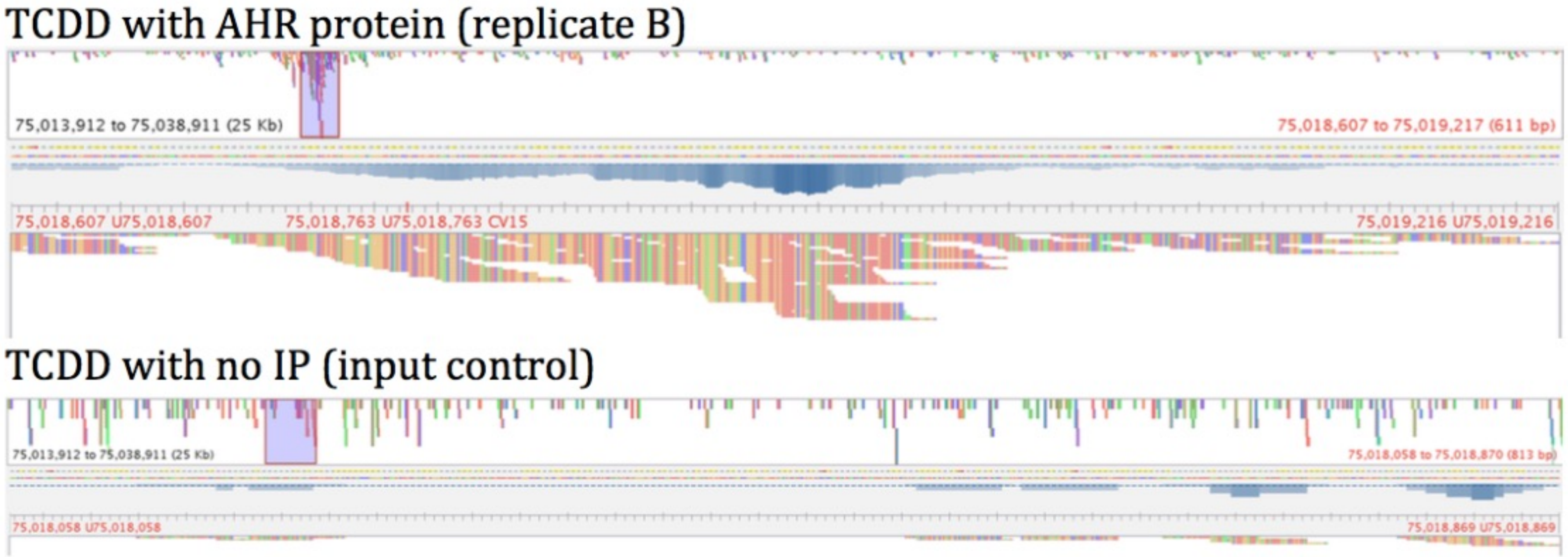
compare

TECHNICAL CONTROL

- DNA extracted from cells
- crosslink DNA-protein
- sonicate
- aliquot soluble chromatin
- ~~immuno-precipitate~~
- reverse formalin crosslinks
- sequence

ChIP-Seq

- NGS reads are mapped to the genome sequence the Burrows-Wheeler Transform
- Peak finding algorithms are used to find small regions in the genome where the sample has more aligned reads than the technical control



TCDD with AHR protein (replicate B)

75,013,912 to 75,038,911 (25 Kb)
75,018,607 to 75,019,217 (611 bp)
75,018,607 U75,018,607
75,018,763 U75,018,763 CV15
75,019,216 U75,019,216

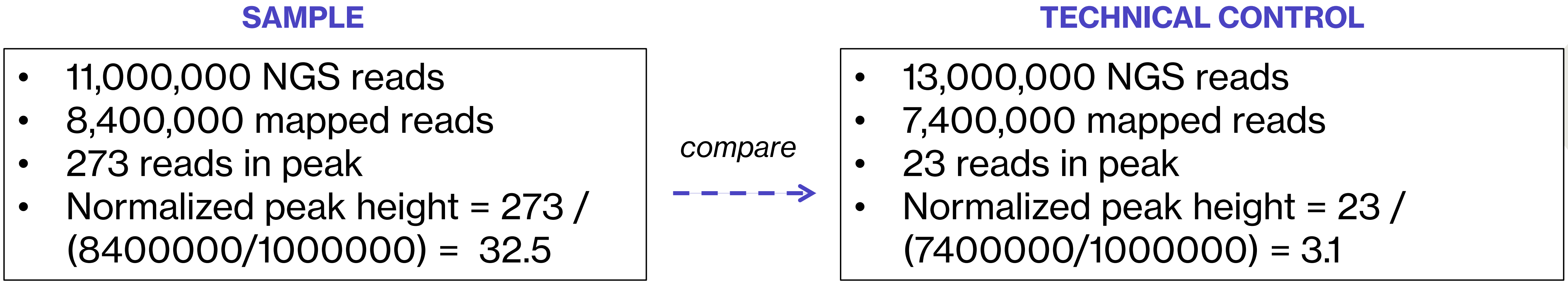
TCDD with no IP (input control)

75,013,912 to 75,038,911 (25 Kb)
75,018,058 to 75,018,870 (813 bp)
75,018,058 U75,018,058
75,018,869 U75,018,869

ChIP-Seq

- For each region found in each sample or input control, perform linear normalization = # reads in region per million mapped reads

- Determine fold enrichment relative to technical control, e.g. for the peak found upstream of CYP1A1



SAMPLE

- 11,000,000 NGS reads
- 8,400,000 mapped reads
- 273 reads in peak
- Normalized peak height = 273 / (8400000/1000000) = 32.5

compare

TECHNICAL CONTROL

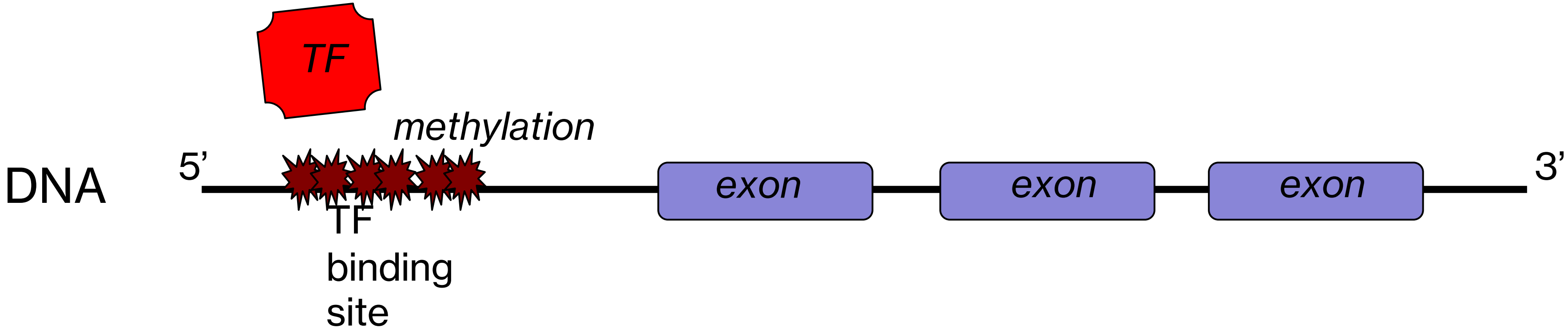
- 13,000,000 NGS reads
- 7,400,000 mapped reads
- 23 reads in peak
- Normalized peak height = 23 / (7400000/1000000) = 3.1

32.5 / 3.1 = 10.45 fold enrichment

ChIP-Seq

- Most ChIP-Seq papers are exploratory – where are the most enriched regions in the genome?
- Reduced NGS sequencing costs leading to replicated experiments
  - Statistical analysis of difference in fold enrichment among sites
  - Familiar false discovery rate issues
- Enriched sites of interests undergo secondary analysis
  - What genes are they upstream of? Does this help elucidate the regulatory role of the target protein?
  - What are the sequences of the putative binding sites? Can we build a PSSM to search for more such binding sites in genomes?
  - Experimental validation by site-directed mutation

Bisulfite-Seq



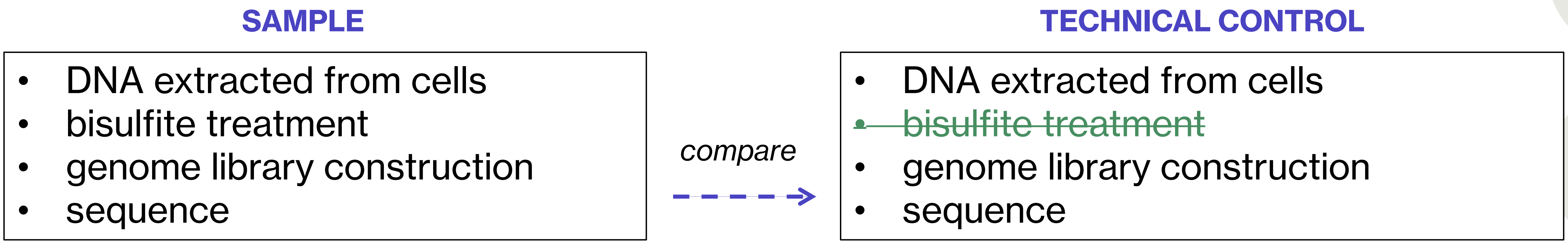
- The goal of this method is to detected methylated regions in the genome

- Combining these data with RNA-Seq and ChIP-Seq among tissues or experimental conditions can lead to a comprehensive understanding of gene regulation

- Whole genome bisulfite sequencing (WGBS) to assess differences in genome-wide cytosine methylation patterns, including CpG, CHH, & CHG

Bisulfite-Seq

- Bisulfite-Seq is essentially a whole genome shotgun sequencing method, i.e. sampling Illumina mate-pairs throughout the genome
- However, bisulfite conversion during library preparation changes unmethylated cytosines to uracil



SAMPLE

- DNA extracted from cells
- bisulfite treatment
- genome library construction
- sequence

compare

TECHNICAL CONTROL

- DNA extracted from cells
- ~~bisulfite treatment~~
- genome library construction
- sequence

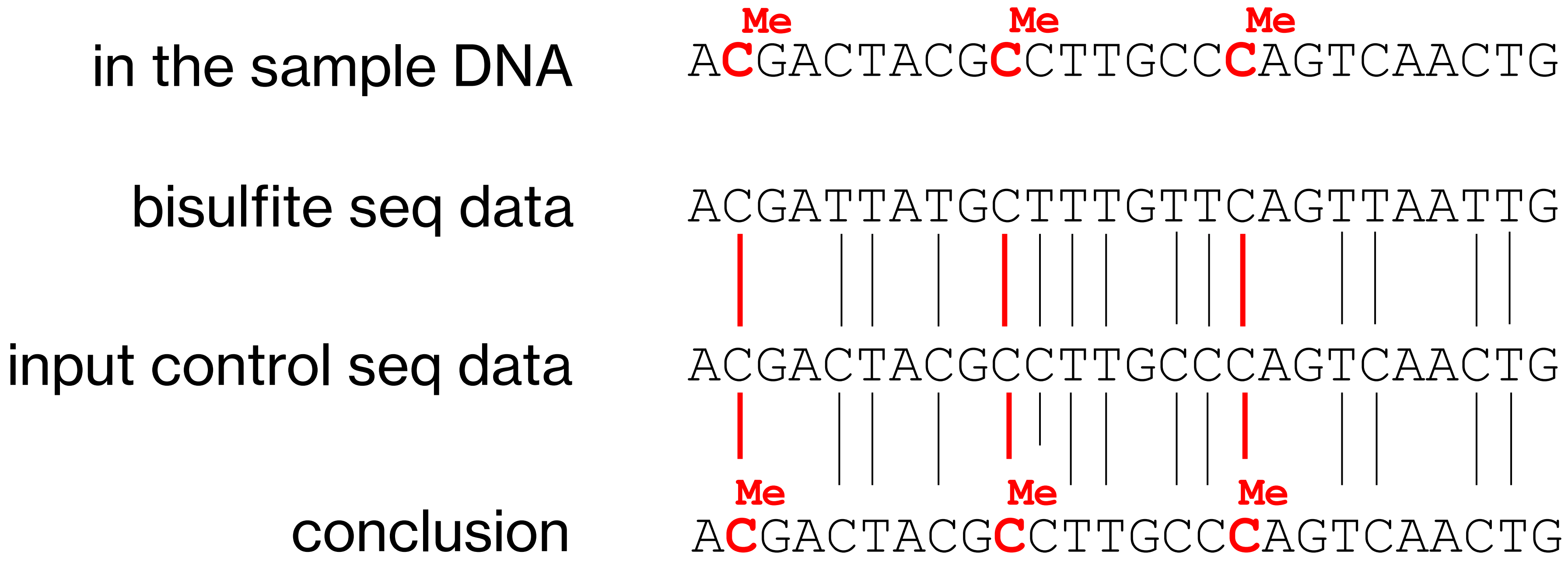
- methylated cytosines: the sample will have cytosines (C) and the technical control will have cytosines (C)
- unmethylated cytosines: the sample will have thymines (T) while the technical control will have cytosines (C)

Bisulfite-Seq

in the sample DNA	$\texttt{A}\overset{\text{Me}}{\texttt{C}}\texttt{GACTACG}\overset{\text{Me}}{\texttt{C}}\texttt{CTTGCC}\overset{\text{Me}}{\texttt{C}}\texttt{AGTCAACTG}$
bisulfite seq data	ACGATTATGCTTTGTTCAGTTAATTG
input control seq data	ACGACTACGCCTTGCCCAGTCAACTG

methylated cytosines: sample will have cytosines (C) and the technical control will have cytosines (C)
unmethylated cytosines: sample will have thymines (T) while the technical control will have cytosines (C)

Bisulfite-Seq



in the sample DNA

Me Me Me

ACGACTACGCCTTGCCCAGTCAACTG

bisulfite seq data

ACGATTATGCTTTGTTCAGTTAATTG

input control seq data

ACGACTACGCCTTGCCCAGTCAACTG

conclusion

Me Me Me

ACGACTACGCCTTGCCCAGTCAACTG

methylated cytosines: sample will have cytosines (C) and the technical control will have cytosines (C)
unmethylated cytosines: sample will have thymines (T) while the technical control will have cytosines (C)

Bisulfite-Seq

- NGS reads are mapped to the genome sequence the Burrows-Wheeler Transform
- DNA methylation levels inferred (based on frequency of Cs or Ts aligned to Cs in the genome) using Bis-SNP variant callers
- Depth of coverage is critical to confidently predict methylation sites
- Whole genome bisulfite sequencing (WGBS) is still somewhat expensive given the costs of sequencing an animal genome, but expected to become more prevalent as NGS costs drop
  - For example, McMaster mouse WGBS of a sample + technical control = ~$3000
- Reduced representation bisulfite sequencing (RRBS) is popular
  - MspI restriction enzyme is used during library construction to enrich for the areas of the genome that have a high CpG content
  - CpG biased sampling of the genome requires less sequencing

Conclusions

- Inventive molecular biology during NGS library construction allows DNA sequencing to be used as an assay for different aspects of biology

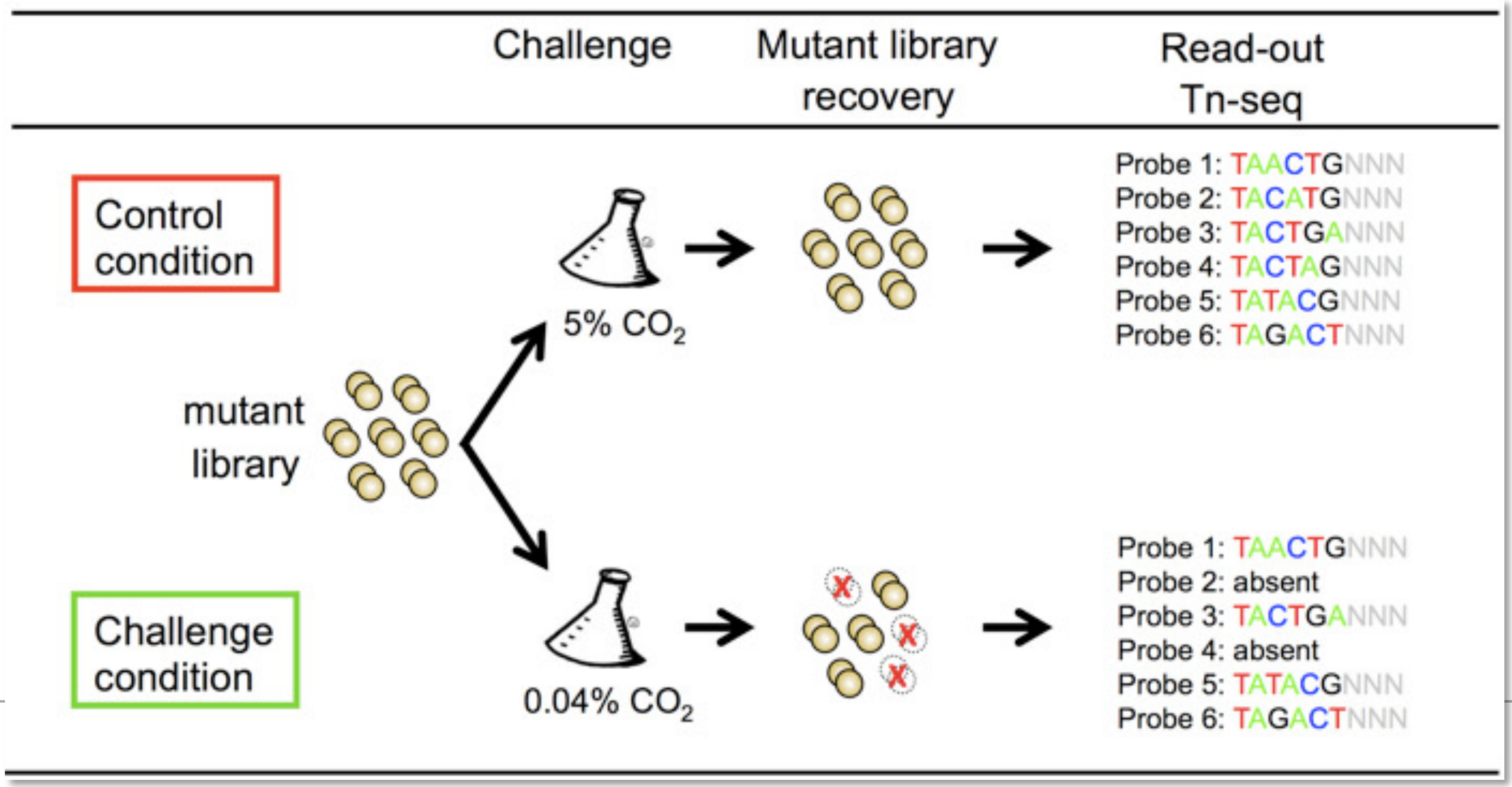
- If you have a method to isolate or enrich a sample for a specific target DNA or RNA population, it can be adapted for NGS analysis at the whole-genome level

- Normalization and false discovery rate issues are generally the same throughout the methods

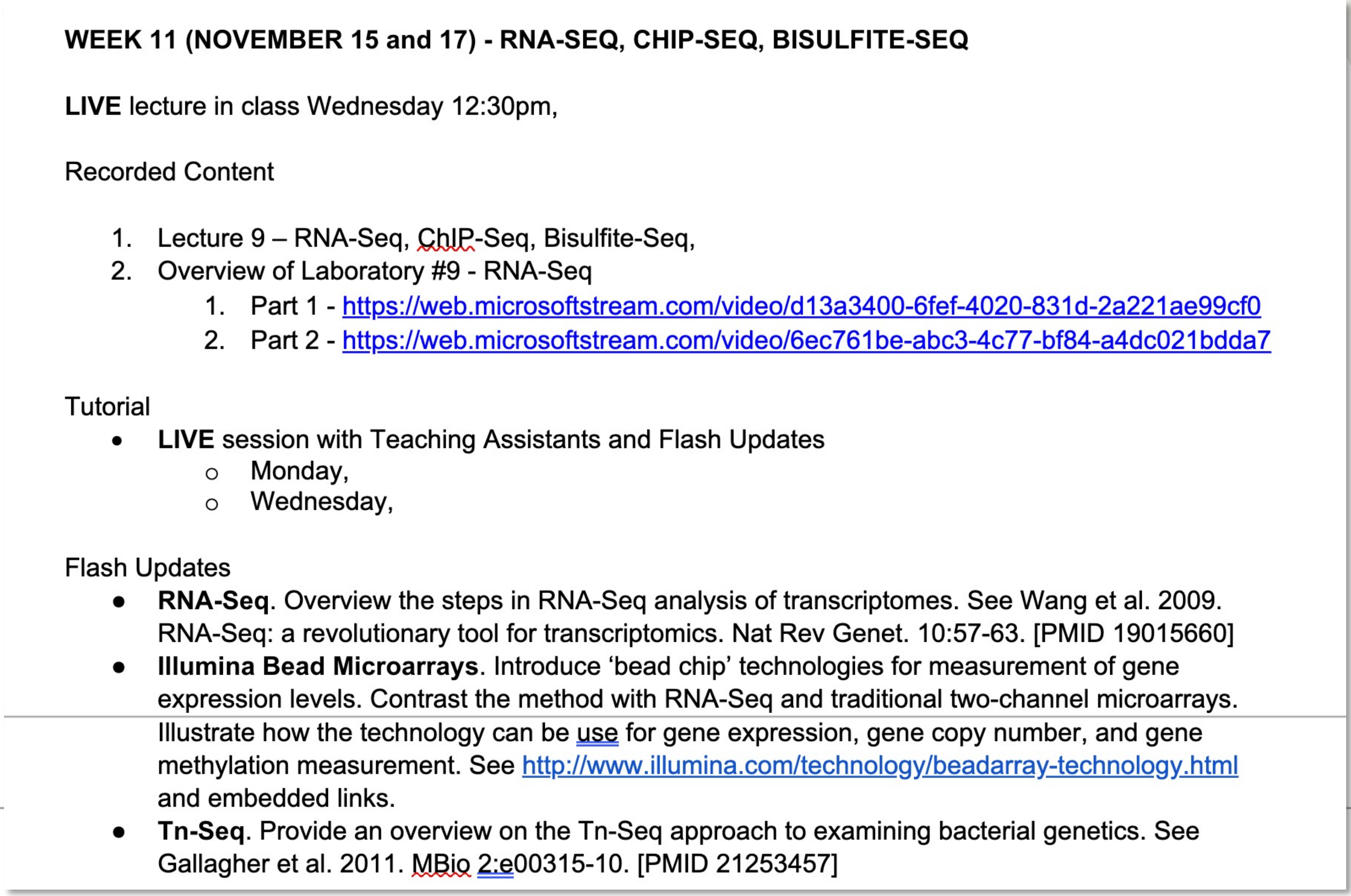
- Use of novel technical controls requires novel algorithm development, e.g. peaking finding in ChIP-Seq, Bis-SNP variant callers for methylation patterns

Conclusions

- Expect to see lots of new assays develop over the next few years, e.g transposon sequencing (Tn-seq) to assay bacteria fitness, gene interactions, gene function



This week...



WEEK 11 (NOVEMBER 15 and 17) - RNA-SEQ, CHIP-SEQ, BISULFITE-SEQ

LIVE lecture in class Wednesday 12:30pm,

Recorded Content

1. Lecture 9 – RNA-Seq, ChIP-Seq, Bisulfite-Seq,
2. Overview of Laboratory #9 - RNA-Seq
    1. Part 1 - https://web.microsoftstream.com/video/d13a3400-6fef-4020-831d-2a221ae99cf0
    2. Part 2 - https://web.microsoftstream.com/video/6ec761be-abc3-4c77-bf84-a4dc021bdda7

Tutorial

- **LIVE** session with Teaching Assistants and Flash Updates
    - Monday,
    - Wednesday,

Flash Updates

- **RNA-Seq**. Overview the steps in RNA-Seq analysis of transcriptomes. See Wang et al. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10:57-63. [PMID 19015660]
- **Illumina Bead Microarrays**. Introduce 'bead chip' technologies for measurement of gene expression levels. Contrast the method with RNA-Seq and traditional two-channel microarrays. Illustrate how the technology can be use for gene expression, gene copy number, and gene methylation measurement. See http://www.illumina.com/technology/beadarray-technology.html and embedded links.
- **Tn-Seq**. Provide an overview on the Tn-Seq approach to examining bacterial genetics. See Gallagher et al. 2011. MBio 2:e00315-10. [PMID 21253457]