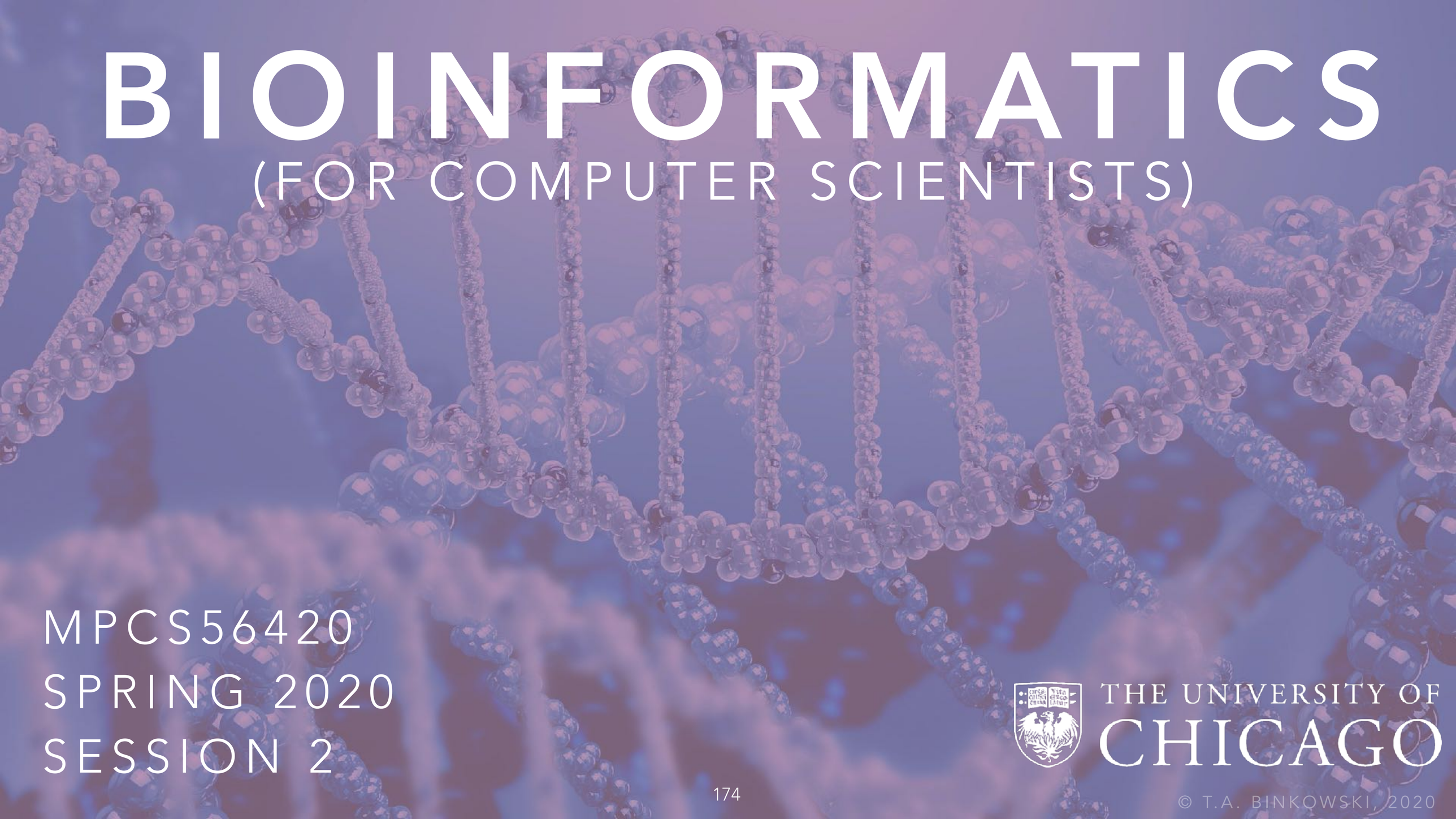


BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)

MPCS56420
SPRING 2020
SESSION 2

THE UNIVERSITY OF CHICAGO

© T.A. BINKOWSKI, 2020


APPLICATIONS
OF SEQUENCING

APPLICATIONS OF SEQUENCING

- The ability to generate large amounts of nucleotide sequence data has revolutionized biology in the past decade
  - Driven by technological advancements
  - New experiments are able to be conducted which were not feasible

APPLICATIONS OF SEQUENCING

- Human microbiome project requirements using Sanger sequencing
  - $667B

SEQUENCE ALL THE BACTERIA IN YOUR GUT

Human Gut Microbiome		Sanger
Number of Species	1000	
Average Genome Size	3 Mb	
Microbiome Size	3 Gb	
Desired Coverage	300 x	
Amount of Data Needed	~ 1 Tb	
Read Length		750bp
Number of **Runs** Needed		~14M
Cost		$667B

APPLICATIONS OF SEQUENCING

- Not only cost, but also time
  - Sanger
  - Roche (next gen)
  - Illumina (next gen)
  - Many competing platforms in NGS

Platform	Reads	Read Length (bases)	Paired Ends	Run Time (days)	Yield (Gb)	Rate (days/Gb)
Sanger	96	750	No	0.5	0.00007	~7000 days
Roche 454 FLX Ti	1 M	400	yes	1	0.8	~1.25 days
Illumina HiSeq 2000	1 B	150	yes	11	300	~1 hr

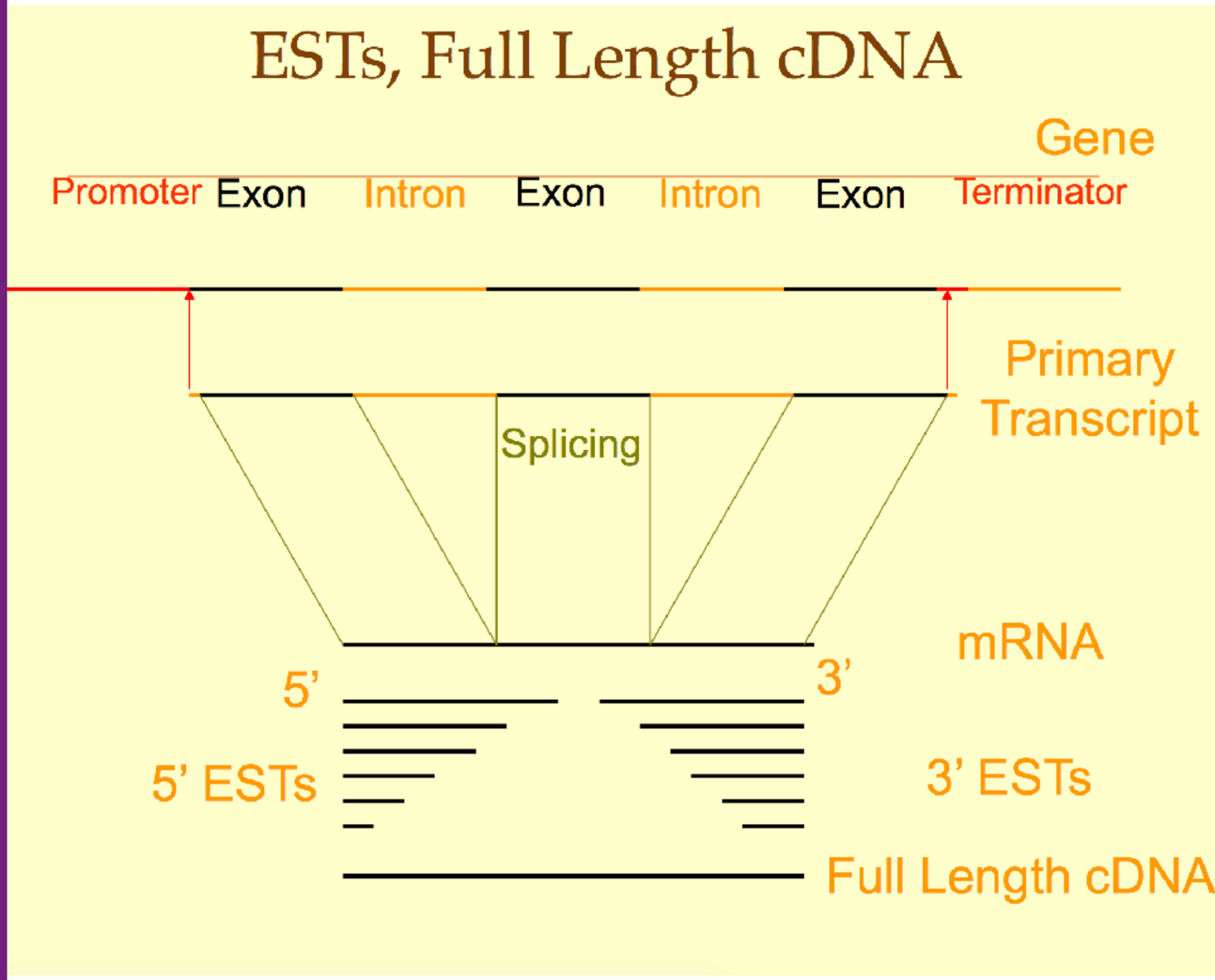
APPLICATIONS OF SEQUENCING

- Next generation sequencing applications
  - de novo sequencing of genomes
  - transcriptomes
  - metagenomes
  - protein-genome interactions

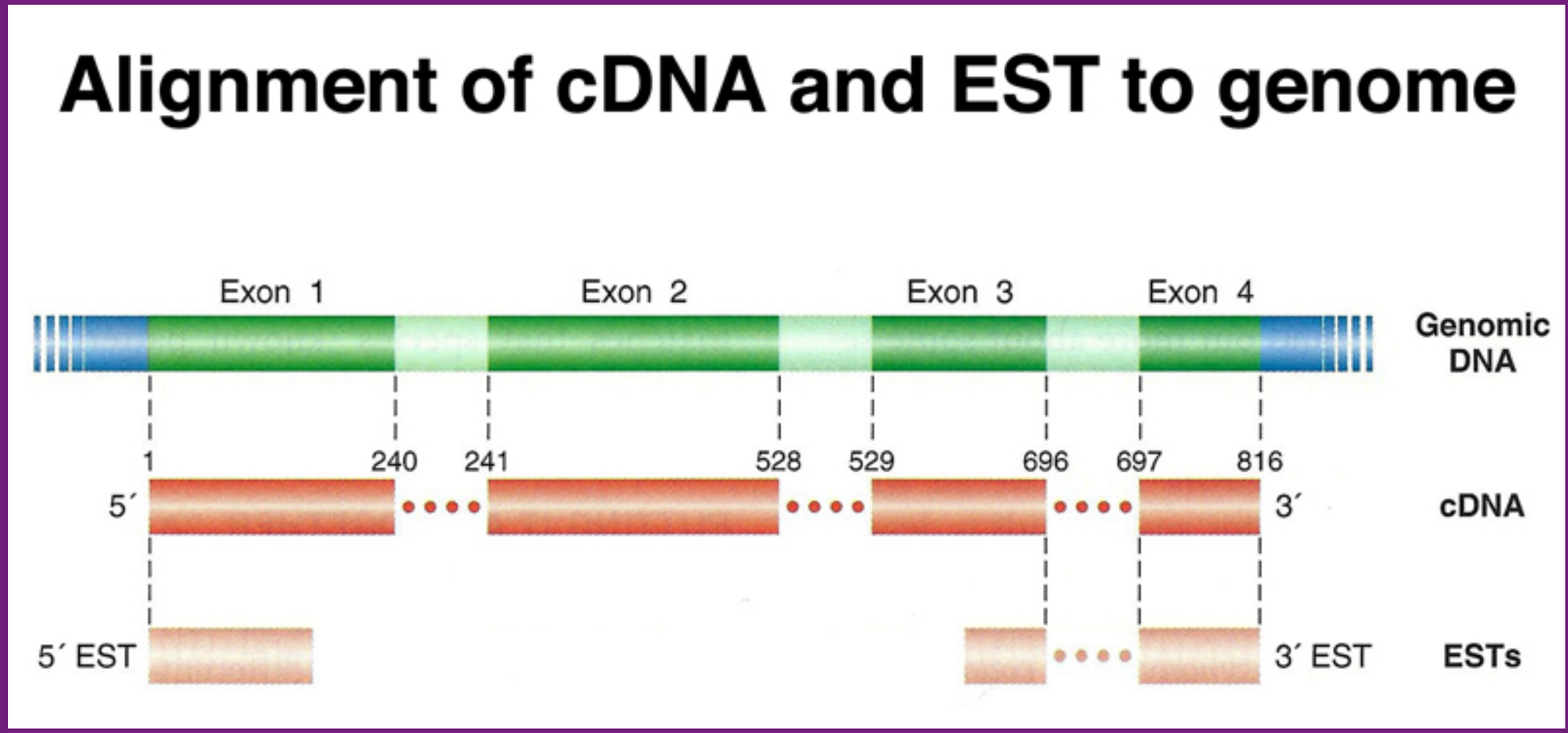
Expressed Sequence Tags

EXPRESSED SEQUENCE TAGS

- ESTs
  - A unique stretch of DNA within a coding region of a gene
  - Short sequences from 5’ or 3’ from mRNA
  - Read: 500-1000 basepairs
  - Used to identify full-length genes
  - Landmark for mapping
  - Result of large scale sequencing of cDNA



EXPRESSED SEQUENCE TAGS



- An identifiable portion of an expressed gene

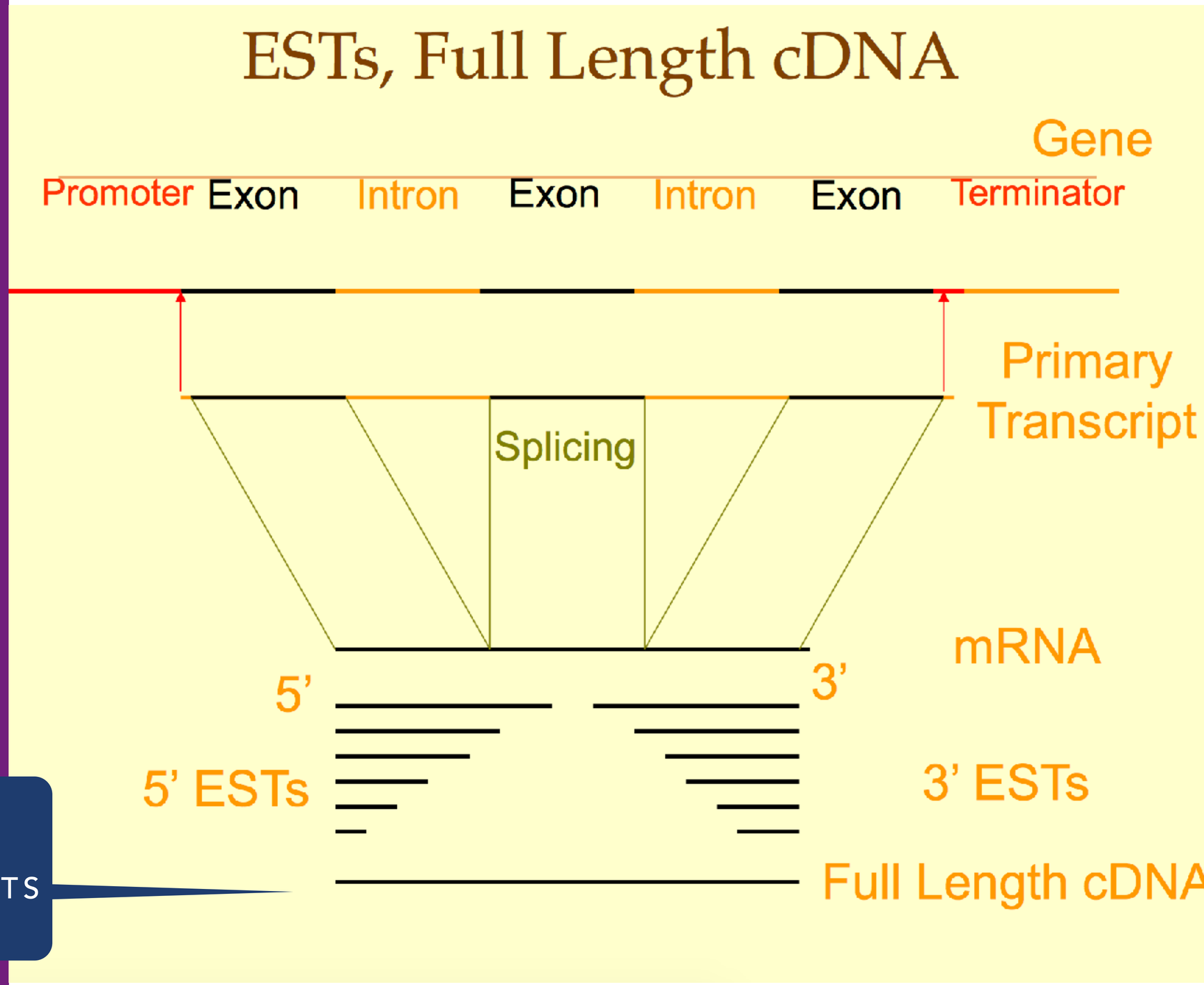
EXPRESSED SEQUENCE TAGS

- Advantages of EST
  - Fast & cheap
  - They represent the most extensive available survey of the transcribed portion of genomes
  - There are indispensable for gene structure prediction, gene discovery and genome mapping:
    - Provide experimental evidence for the position of exons
  - Characterization of splice variants (e.g. different tissues)
  - Sequences of multiple ESTs can reconstitute a full-length cDNA

EXPRESSED SEQUENCE TAGS

- NCBI dbEST
- UniGene - Cluster ESTs

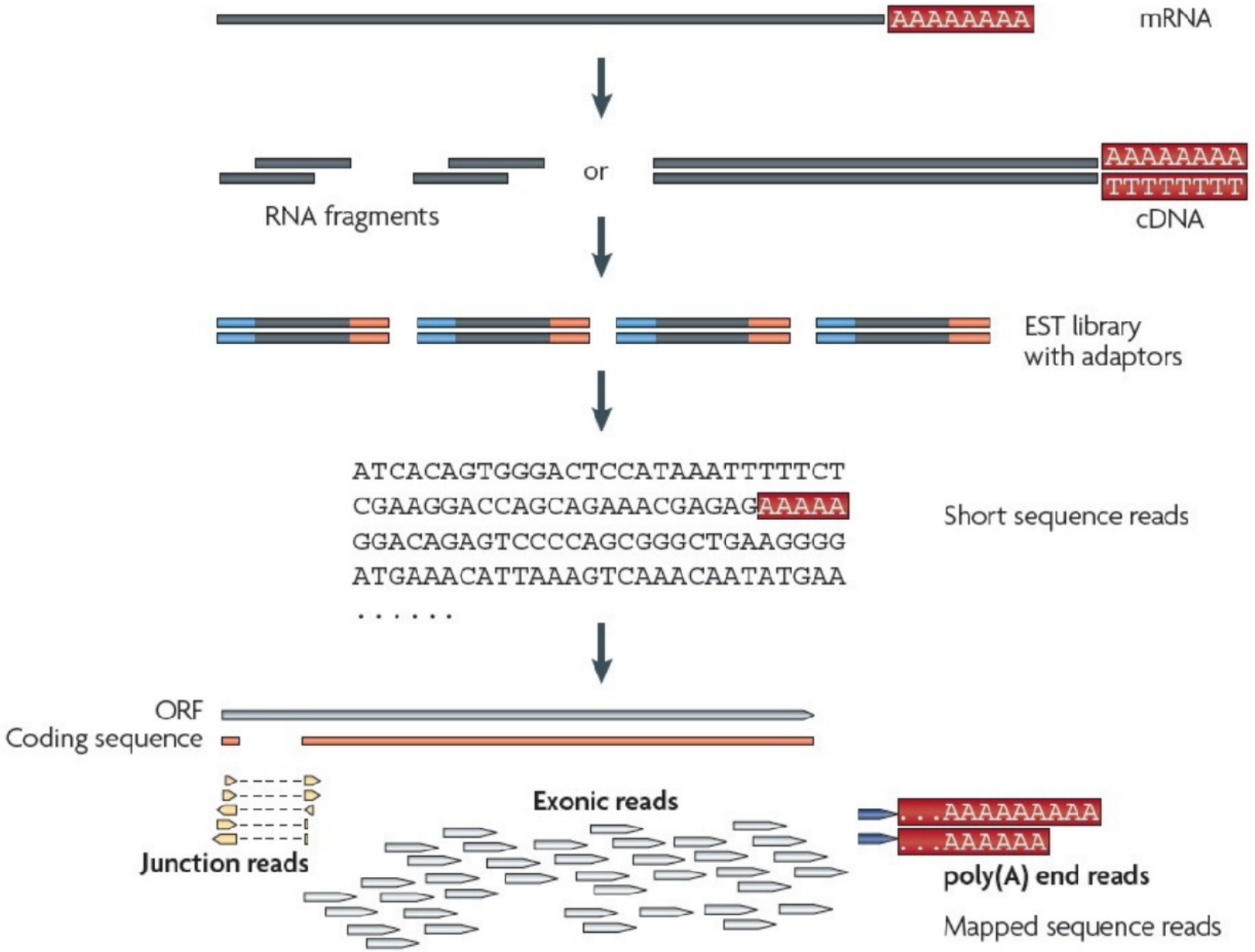
FULL LENGTH CDNA FROM ESTS



RNA-Seq

RNA-SEQ

- RNA-Seq (Whole Transcriptome Sequencing)
- Massively parallel sequencing method for transcriptome analyses

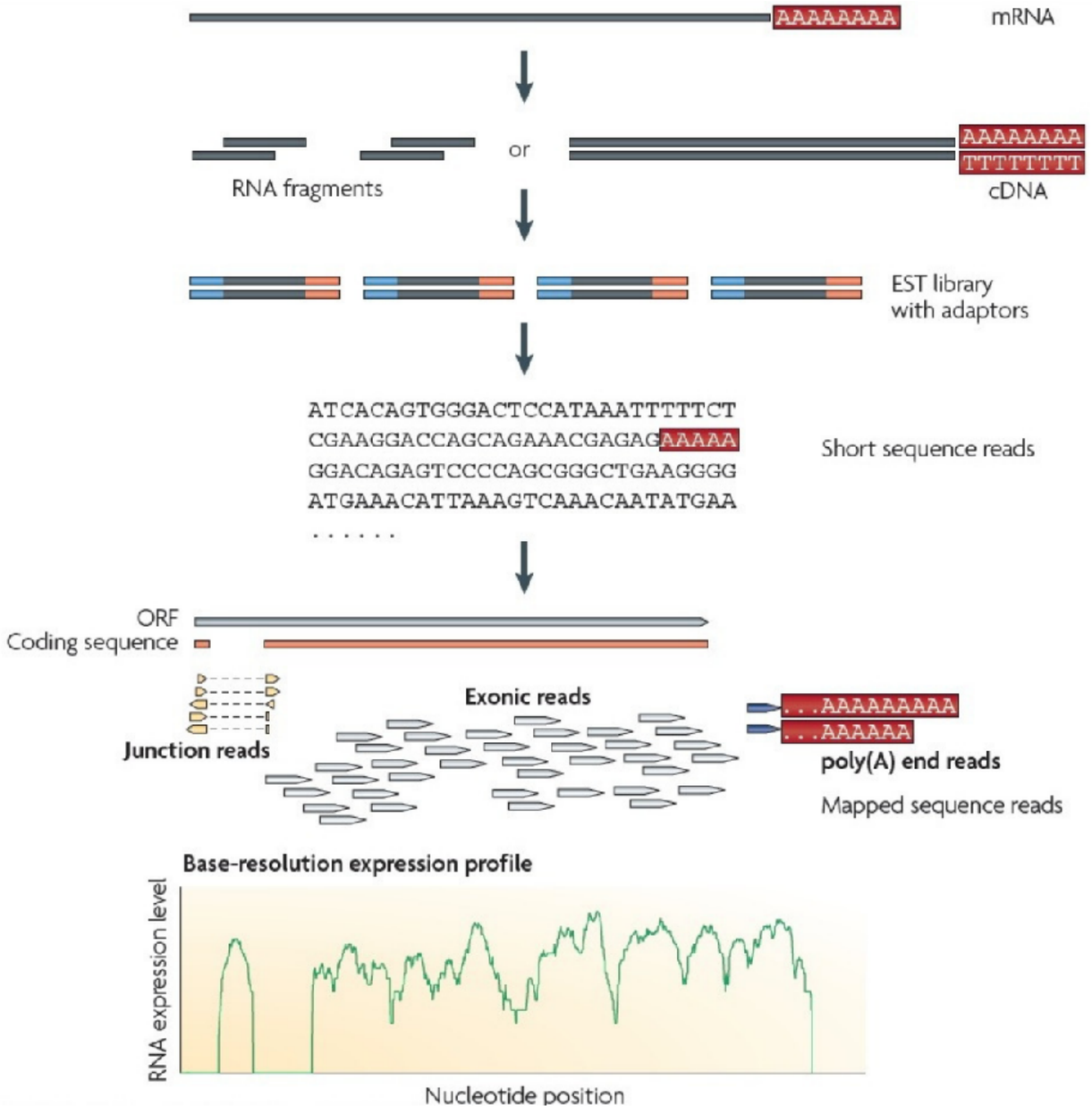


RNA-SEQ

- Aims of RNA-seq
  - To quantify mRNA abundance
  - To determine the transcriptional structure of genes: start sites, 5′ and 3′ ends, splicing patterns
  - To quantify the changing expression levels of each transcript during development and under different conditions
- Identify novel sequences

RNA-SEQ

- Method overview
  - Complementary DNA (cDNA) generated from RNA are sequenced using next-generation “short read” technologies
  - Reads are aligned to a reference genome and a transcriptome map is constructed
- Evaluation of alternative splicing events may be visualized in a genome browser

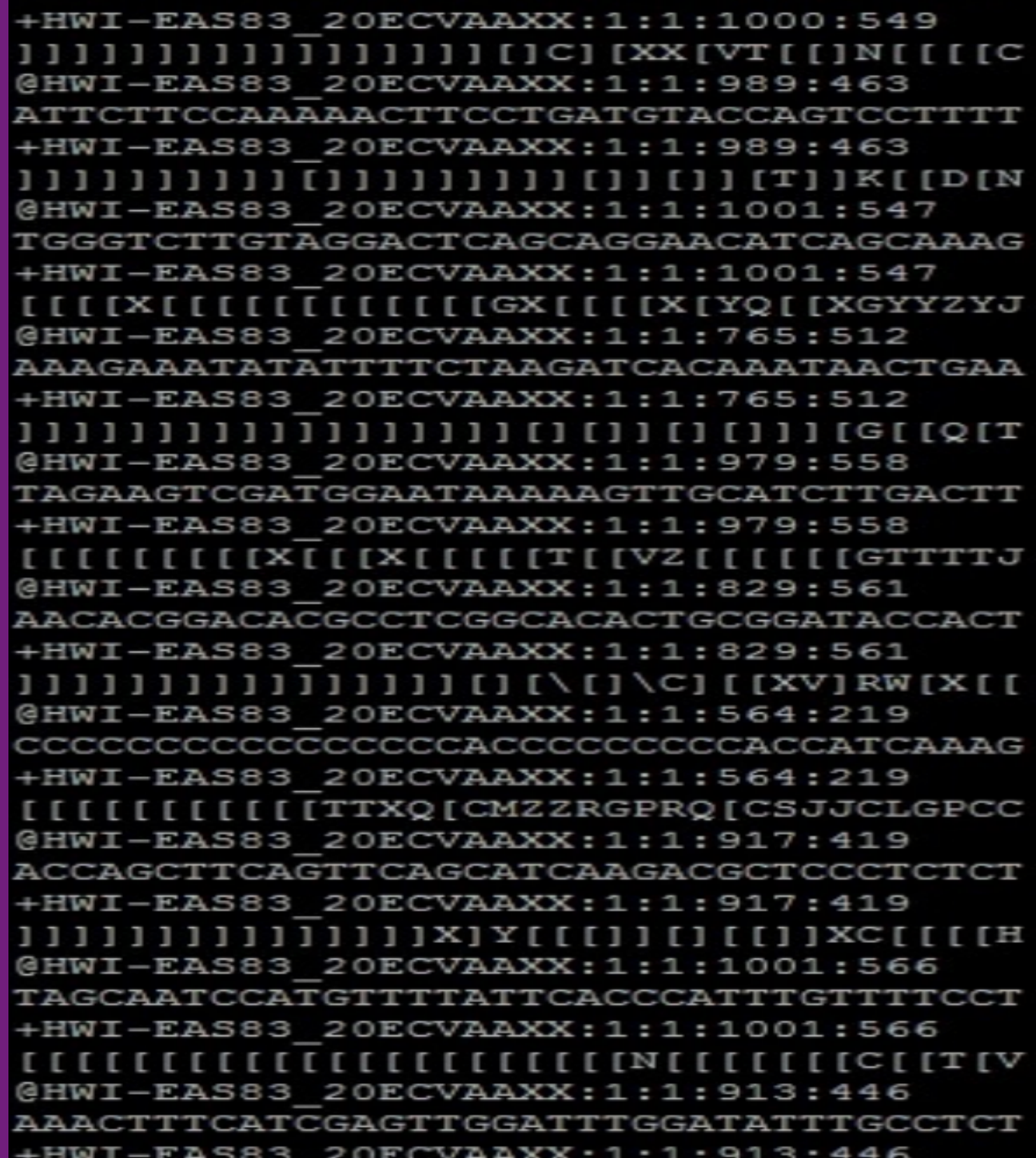


RNA-SEQ

- Transcriptome
  - The complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition
    - Tissue specific
    - Disease progression
  - Interpreting the functional elements of the genome
  - Revealing the molecular constituents of cells, tissues
  - Understanding development and disease

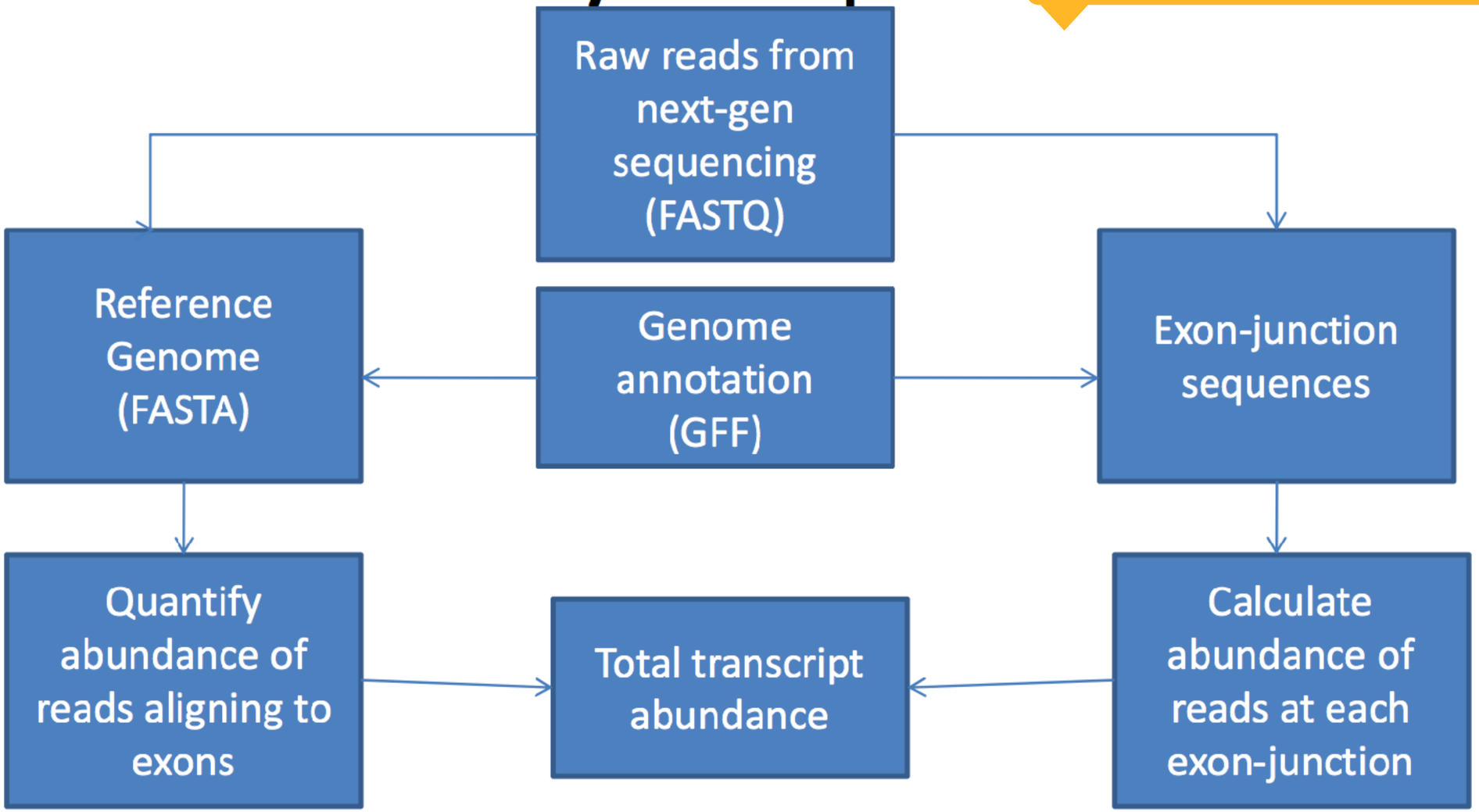
RNA-SEQ

- Technology
  - Single-end, paired-end
  - Typically 30-400bp reads
  - Popular platforms:
    - Illumina, 454, SOLID
    - 10 million reads
  - Alignment tools
    - Bowtie, BWA, Eland etc
  - Additional step: align to exon-junctions
  - Automated pipeline for RNA-Seq:
    - Tophat : for alignment
    - Cufflinks : for calculating expression levels



RNA-SEQ

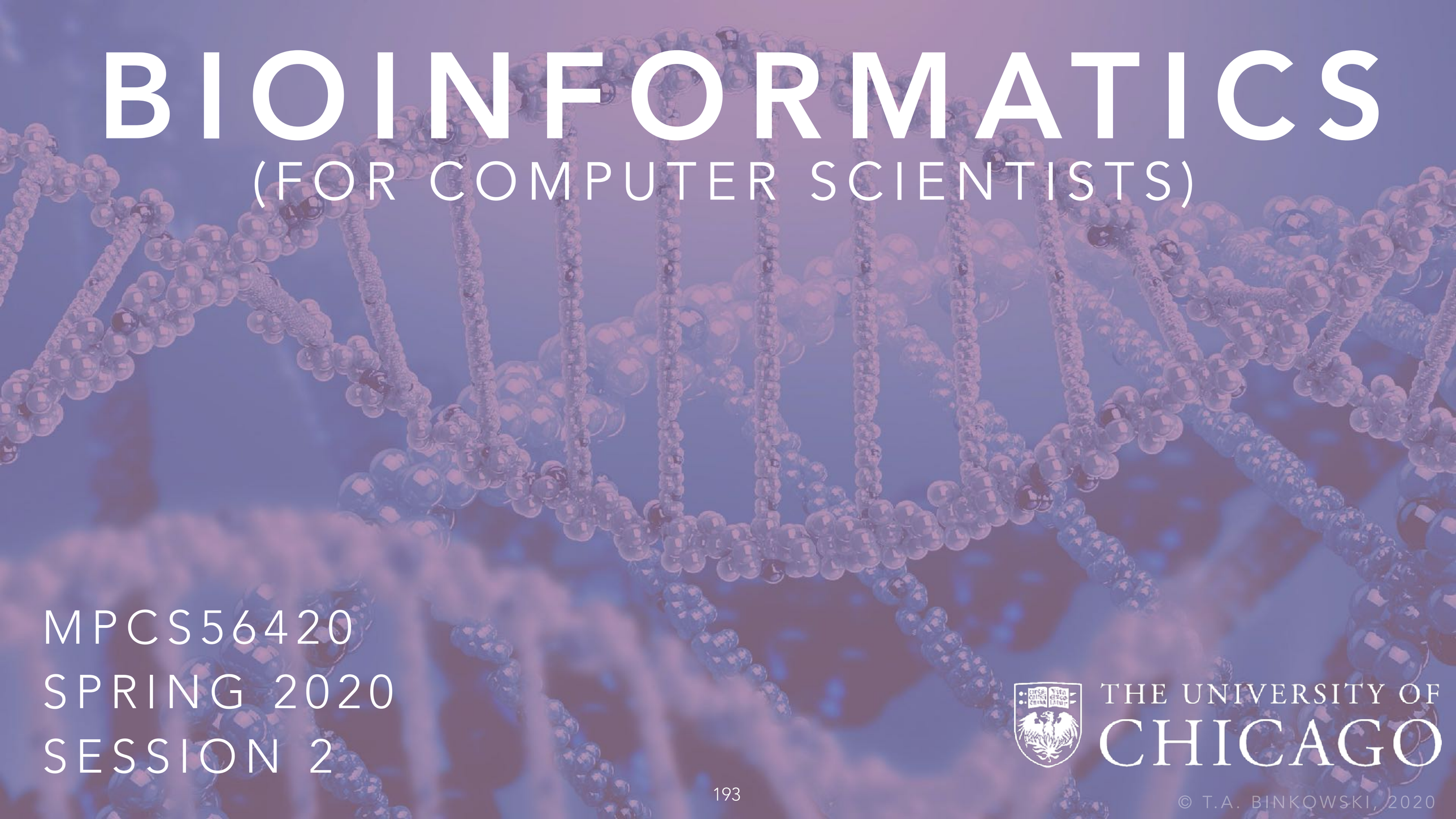
ANALYSIS PIPELINE



RNA-SEQ

ADVANTAGE OF RNA-SEQ

- Does not require existing genomic sequence
- Very low background noise
  - Reads can be unambiguously mapped
- High-resolution
- High-throughput
  - Better than Sanger sequencing of cDNA or EST libraries
- Cost
  - Lower than traditional sequencing
- Can reveal sequence variations (SNPs)



BIOINFORMATICS

(FOR COMPUTER SCIENTISTS)

MPCS56420
SPRING 2020
SESSION 2

THE UNIVERSITY OF CHICAGO

© T.A. BINKOWSKI, 2020