GMS6850 – Core Concepts in Bioinformatics
Lecture 2
Translation, sequence alignment, short read mapping

2021 01 13

Steve Rozen

Translation: example worked on board

Sequence alignment (concept map)

Pairwise alignment

Multiple alignment

Global alignment (Needleman-Wunsch)

Local alignment (Smith-Waterman)

Scoring matrices

Dot plots

RNA secondary structure

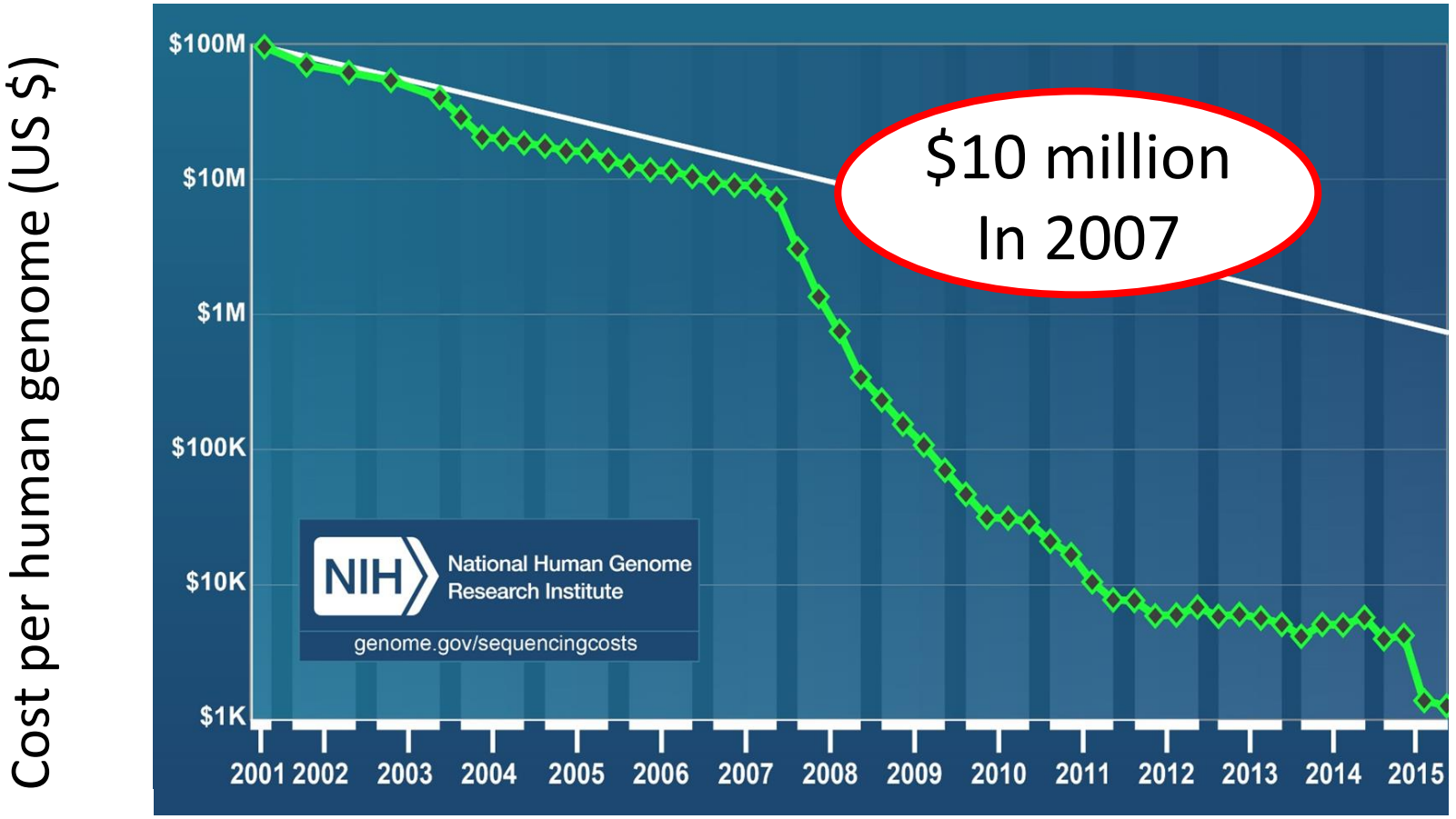
Short read mapping / alignment in next generation sequencing

Sequence database search (for example BLAST) for a later class

Slides from alignment folder

Short read technology and short read mapping

NGS has been a disruptive technology that underpins many scientific opportunities

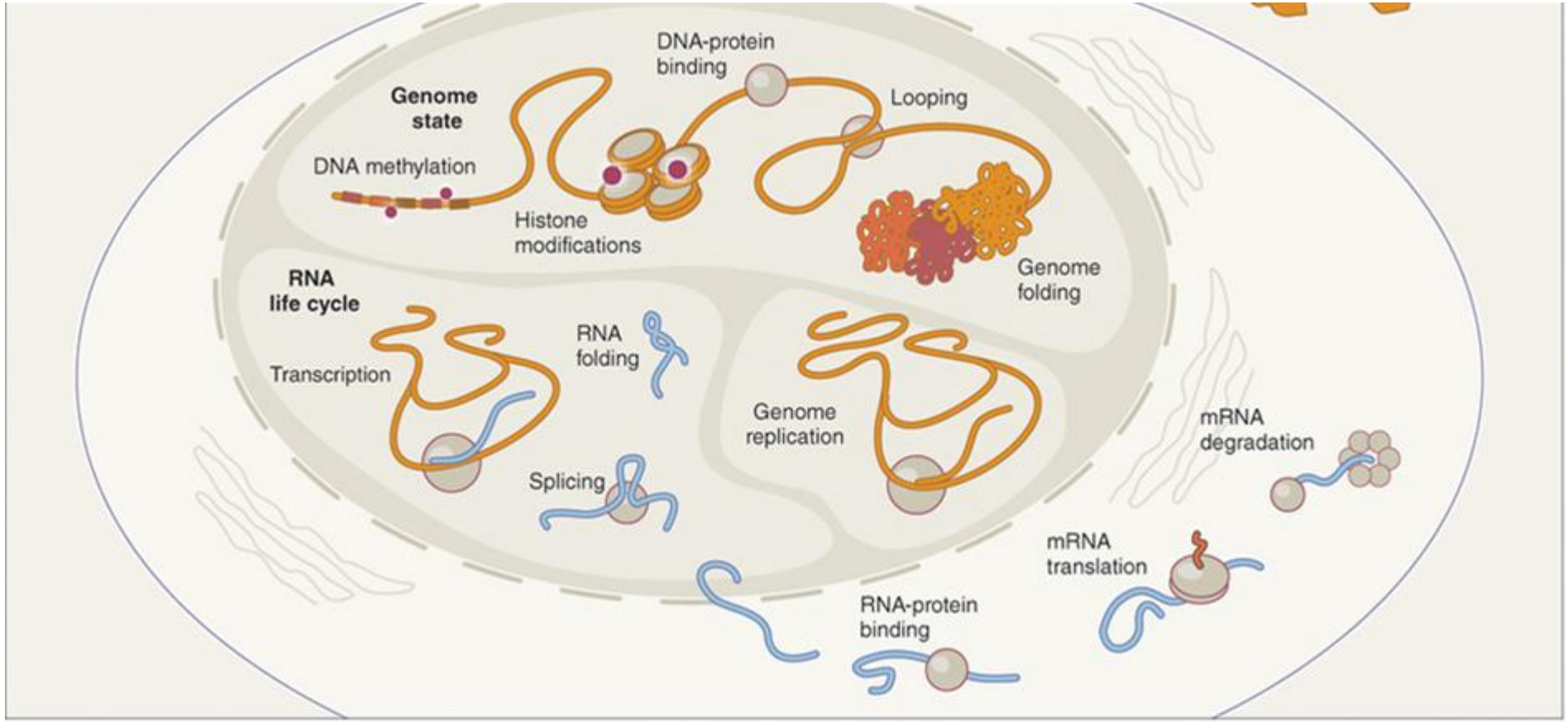


http://www.genome.gov/sequencingcosts/

6,000 X
drop in price

$700
today

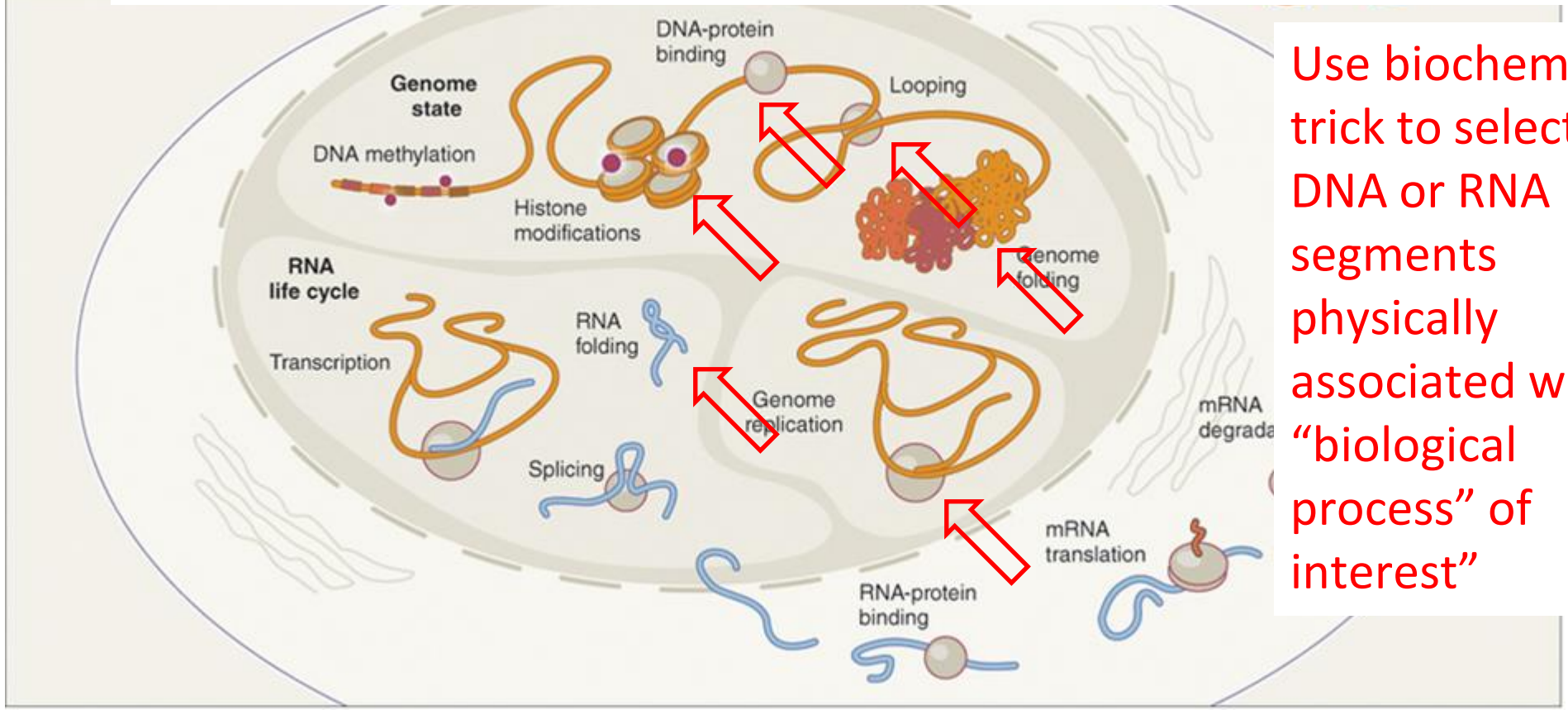
Next generation sequencing (NGS) is a foundational technology



Shendure & Aiden, Nat. Biotech. 2012

Next generation sequencing (NGS) will become a foundational technology

Snippets of DNA or RNA are self-identifying (self-barcoded)



Use biochemical trick to select for DNA or RNA segments physically associated with "biological process" of interest"

Applications of NGS

Method	To determine…
DNA-Seq	A genome sequence
Targeted DNA-Seq	A subset of a genome (for example, an exome)
RNA-Seq	RNA (that is, the transcriptome)
Methyl-Seq	Sites of DNA methylation, genome-wide
Targeted methyl-Seq	DNA methylation in a subset of the genome
DNase-Seq, Sono-Seq	Active regulatory chromatin (nucleosome-depleted)
FAIRE-Seq (formaldehyde-assisted isolation of regulatory elements)	Active regulatory chromatin (nucleosome-depleted)
MAINE-Seq (MNase-assisted isolation of nucleosomes)	Histone-bound DNA (nucleosome positioning)
ChIP-Seq	Protein-DNA interactions (using chromatin immunoprecipitation)
RIP-Seq (RNA-binding protein immunoprecipitation)	Protein-RNA interactions
CLIP-Seq (cross-linking IP)	Protein-RNA interactions

Adapted from Shendure and Aiden, Nat. Biotech. 2012

Applications of NGS

Method	
DNA-Seq	
Targeted DNA-Seq	
RNA-Seq	
Methyl-Seq	
Targeted methyl-Seq	
DNase-Seq, Sono-Seq	
FAIRE-Seq (formaldeh isolation of regulatory	
MAINE-Seq (MNase-a nucleosomes)	
ChIP-Seq	
RIP-Seq (RNA-binding protein immunoprecipitation)	Protein-RNA interactions
CLIP-Seq (cross-linking IP)	Protein-RNA interactions

Many applications

- Human variation / human genetics
- Cancer genetics
- Clinical applications in human genetics and cancer genetics
- Plant and animal breeding for agriculture
- Sequencing new genomes
- Metagenomics
- Pathogen discovery
- Pathogen evolution

Adapted from Shendure and Aiden, Nat. Biotech. 2012

Sequencing technology

Dominant technology – short reads

DNBSEQ (BGI) technology – different chemistry but plug compatible with Illumina

https://www.bgi.com/us/resources/sequencing-platforms/

Ion Torrent – different chemistry NOT plug compatible with Illumina

https://www.thermofisher.com/sg/en/home/brands/ion-torrent.html

Illumina high throughput sequencers

	 HiSeq X Five†	 HiSeq X Ten†	 NovaSeq 6000
Output Range	900-1800 Gb	900-1800 Gb	134-6000 Gb‡
Run Time	< 3 days	< 3 days	24-44 hr
Reads per Run	3-6 billion	3-6 billion	Up to 20 billion‡
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp
Samples per Run§	8-16	8-16	4-48
Relative Price per Sample§	Higher Cost	Lower Cost	Higher Cost
Relative Instrument Price§	Lower Cost	Higher Cost	Lower Cost

Review of Illumina Sequencing Technology (document at URL below)

https://sapac.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Other sequencing technology, long reads

https://nanoporetech.com/

https://www.pacb.com/

Sequence analysis pipelines and short read mapping

Whole-genome Analysis Pipeline

Sequencing Provider

Illumina Hiseq X
30x coverage
150 bp paired-end

FASTQ file

FASTQC

FASTQ quality checker

FASTQ files (2 X 30 Gb / sample)

BWA-MEM

Short read aligner (used often for genomic data)

Sambamba

Miscellaneous post-processing

BAM files (~70 Gb / sample)

Qualimap

Qualimap
http://qualimap.bioinfo.cipf.es/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4708105/

HLA typing

Virome profiling

Freebayes Joint Calling

Single Nucleotide Variant (SNV) caller

Parliament

Structural (SV) and Copy Number Variant (CNV) caller

Multisample VCF (18 Gb for 100 samples)

SNVs + Indels

SVs + CNVs

Annotation: is this a known variant?
Does it affect protein sequence?

Annotation

Annotation

SPECTRA DB

Database to store variant calls